STATISTICAL METHODS FOR THE DETECTION AND ANALYSES OF STRUCTURAL VARIANTS IN THE HUMAN GENOME

TEO SHU MEI

NATIONAL UNIVERSITY OF SINGAPORE

KAROLINSKA INSTITUTET

2012

STATISTICAL METHODS FOR THE DETECTION AND ANALYSES OF STRUCTURAL VARIANTS IN THE HUMAN GENOME

TEO SHU MEI

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

SAW SWEE HOCK SCHOOL OF PUBLIC HEALTH

NATIONAL UNIVERSITY OF SINGAPORE

DEPARTMENT OF MEDICAL EPIDEMIOLOGY AND BIOSTATISTICS

KAROLINSKA INSTITUTET

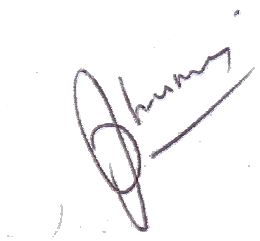
STOCKHOLM, SWEDEN

2012

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



Teo Shu Mei
26 Nov 2012

SUMMARY

Structural variations (SVs) are an important and abundant source of variation in the human genome, encompassing a greater proportion of the genome as compared to single nucleotide polymorphisms (SNPs). This thesis investigates different aspects of SV analysis, focusing on copy number variations (CNVs) and regions of homozygosity (ROHs). It is divided into four main studies, each focusing on a different set of aims.

In Study I, Identification *of recurrent regions of copy-number variation across multiple individuals*, we develop an algorithm and software to identify common CNV regions using individually segmented data. The identified common regions allow us to investigate population characteristics of CNVs, as well as to perform association studies.

In Study II, *Multi-platform segmentation for joint detection of copy number variants*, we develop an algorithm to identify CNVs using intensity data from more than one platform. The algorithm is useful when researchers have data from multiple platforms on the same individual.

In Study III, *Regions of homozygosity in three Southeast-Asian populations*, we identify ROHs in three Singapore populations, namely the Chinese, Malays and Indians. We characterize the regions and provide population summary statistics. We also investigate the relationship between the occurrence of ROHs and haplotype frequency, regional linkage disequilibrium (LD) and positive selection. The results show that frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis, demonstrating that ROHs contain information on population structure and the demographic history of a population.

Last but not least, in Study IV, *Statistical challenges associated with detecting copy number variants with next-generation sequencing technology*, we describe and discuss areas of potential biases in CNV detection for each of four commonly used methods. In particular, we focus on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulties in identifying duplications. To gain insights to some of the issues discussed, we download real data from the 1000 Genomes Project and analyze it in terms of depth of coverage (DOC). We show examples of how reads in repeated regions can affect CNV detection, demonstrate current GC correction algorithms, investigate sensitivity of DOC algorithm before and after quality-control of reads and discuss reasons for which duplications are harder to detect than deletions.

PREFACE

I first started dabbling with genetic data during my 4th year as a Statistics undergraduate in 2007. I was working on the Affymetrix 500K SNP array, one of the densest SNP microarrays at that time. Barely 5 years later, there are arrays with more than 5 million SNPs, not to mention Next-generation sequencing arrays that produce billions of reads in a single run. The technologies to study genetics have certainly evolved very rapidly, bringing with it new challenges in terms of statistical and bioinformatics analyses.

When I first learnt of the term 'CNV', the concept sounded simple to me: That we have regions of the genome that are deleted/duplicated, and that based on the intensity of our measurements, less intense means less of that particular region, and vice versa. "Not too complex!" I thought naively. As I continue to learn more, the multitude of problems/challenges that comes associated with the analysis of noise-rich CNV data is enormous. As put across aptly by John Ioannidis on genetic data from microarrays in general, "…this noise is so data-rich that minimum, subtle, and unconscious manipulation can generate spurious "significant" biological findings that withstand validations by the best scientists, in the best journals. Biomedical science would then be entrenched in some ultramodern middle ages, where tons of noise is accepted as "knowledge". – The Lancet 365: 454-455.

Nevertheless, I hope that with these four years of hard work, I have helped made a little more sense out of the massive amount of genetic data we have.

LIST OF PUBLICATIONS

This thesis is based on the following original articles which will be referred to in the text by their Roman numerals.

I. **Teo SM**, Salim A, Calza S, Ku CS, Chia KS, Pawitan Y. (2010) Identification of recurrent regions of copy number variation across multiple individuals. *BMC Bioinformatics* **11**:147.

II. **Teo SM**, Pawitan Y, Kumar V, Thalamuthu A, Seielstad M, Chia KS, Salim A. (2011) Multi-platform Segmentation for joint detection of copy number variants. *Bioinformatics* **27**:11.

III. **Teo SM**, Ku CS, Salim A, Naidoo N, Chia KS, Pawitan Y. (2012) Regions of homozygosity in three Southeast Asian populations. *Journal of Human Genetics* **57**: 101-108.

IV. **Teo SM**, Pawitan Y, Ku CS, Chia KS, Salim A. Statistical challenges associated with detecting copy number variants with next-generation sequencing technology. *Manuscript.*

Other relevant publications:

- **Teo SM**, Ku CS, Naidoo N, Hall P, Chia KS, Salim A, Pawitan Y. (2011) A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. *Journal of Human Genetics* **56**:524-533.

- Ku CS, **Teo SM**, Naidoo N, Sim X, Teo YY, Pawitan Y, Seielstad M, Chia KS, Salim A. (2011) Copy number polymorphisms in new HapMap III and Singapore populations. *Journal of Human Genetics* **56**:552-560.

- Ku CS, Naidoo N, **Teo SM**, Pawitan Y. (2011) Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* **129**:1-15.

- Ku CS, Naidoo N, **Teo SM**, Pawitan Y. (2011) Characterising Structural Variation by Means of Next-Generation Sequencing. *Encyclopedia of Life Sciences (ELS)*. John Wiley & Sons, Ltd: Chichester.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 5.1: Diagram illustrating the non-triviality of determining if two CNVs are the 'same' variant. In (a), CNV1 and CNV2 overlap completely. In this case, we are confident that the two CNVs are the same. In (b), the start and end positions of CNV1 and CNV2 differs, but there is substantial overlap between the two. In (c), CNV1 is completely within the range of CNV2 but the two CNVs differ vastly in lengths. In most research papers, scientists are comfortable with using a 50% reciprocal overlap to determine if two CNVs are concordant.

LIST OF ABBREVIATIONS

The following abbreviations have been used in this thesis and in the associated four original publications:

aCGH	Array comparative genomic hybridization
AIC	Akaike information criterion
ANOVA	Analysis of variance
AS	Assembly based
BAF	B allele frequency
Bp	Base-pairs
CAHRES	Cancer Hormone Replacement Epidemiology in Sweden
CBS	Circular Binary Segmentation
COMPOSITE	Cumulative Overlap Using Very Reliable Regions
COVER	Cumulative Composite Confidence Scores
CNV	Copy number variation
CNVR	Copy number variation region
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
DOC	Depth of coverage
EHH	Extended haplotype homozygosity
FDR	False discovery rate
GWAS	Genome-wide association studies
HIV	Human immunodeficiency virus
HMM	Hidden Markov model
HTS	High throughput sequencing
HWE	Hardy Weinberg equilibrium
iHS	Integrated haplotype score
kb	Kilo base-pairs
LD	Linkage disequilibrium
LINEs	Long interspersed nuclear elements
LOH	Loss of homozygosity
LRR	Log R ratio

LTR	Long terminal repeat
MAF	Minor allele frequency
MPSS	Multi-platform smooth segmentation
MPCBS	Multiple platform circular binary segmentation
NGS	Next-generation sequencing
PEM	Paired end mapping
PCA	Principal component analysis
PCR	Polymerase chain reaction
QC	Quality-control
RD	Read depth
rDNA	Ribosomal deoxyribonucleic acid
RP	Read pair
ROH	Regions of homozygosity
SINEs	Short interspersed nuclear elements
SOLiD	Supported Oligonucletide Ligation Detection System
SMS	Single molecule sequencing
SNP	Single-nucleotide polymorphism
SR	Split read
SV	Structural variants
TGS	Third generation sequencing
VKORC1	Vitamin K epoxide reductase complex subunit 1
VNTR	Variable number of tandem repeats
WTCCC	Wellcome Trust Case Control Consortium

Chapter 1 – INTRODUCTION

Genetic variation in the human genome can take many forms, including single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), indels, regions of homozygosity (ROHs), and other structural variants (SVs). In the last couple of years, genome-wide association studies (GWAS) have been widely used to correlate genetic differences to phenotypic variation, but they were largely focused on SNPs.

CNVs and other SVs were less appreciated until two landmark studies in 2004 identified widespread deletions and duplications in the human genome (Sebat *et al*., 2004; Iafrate *et al*., 2004). By now, CNVs are widely recognized as a prevalent form of variation in the genome, encompassing a greater proportion of the genome as compared to SNPs. An estimated 1.2% of a single genome differs from the reference human genome when considering CNVs, as compared to 0.1% by SNPs (Pang *et al*., 2010). Recent studies have found CNVs to be associated with complex diseases such as human immunodeficiency virus (HIV) infection, cancer, diabetes, mental disorders, obesity, Parkinson's disease and autoimmune diseases (Wain *et al*., 2009; The Wellcome Trust Case Control Consortium 2010). ROHs are also more abundant than previously thought (Gibson *et al*., 2006), and are associated with complex diseases such as schizophrenia and late-onset Alzheimer's disease (Lencz *et al*., 2007; Nalls *et al.*, 2009).

That, as compared to SNPs, the association of CNVs and ROHs with complex diseases is not as well-studied is in part due to greater complexity in identifying these multi-base, multi-allelic variants, and also greater complexity in performing

association studies with these variants. Early works on CNVs/ROHs have focused largely on identifying and characterizing regions in the genome which harbour them. This has been necessary in laying the foundation to improve our understanding of CNVs/ROHs for subsequent association analysis with human complex diseases.

The most common technologies for CNVs identification in the last couple of years are high density SNP arrays and array comparative genomic hybridization (aCGH) arrays; the former (SNP arrays) are also commonly used for detection of ROHs. However, the data generated from these techniques are noisy, and identifying CNVs comprehensively with high resolution still remains a technical and statistical challenge. aCGH and SNP arrays are also limited by the resolution of the array to determine precise locations of CNV breakpoints, and are unable to locate copy-neutral events such as inversions and translocations.

Sanger sequencing, often seen as the gold standard for CNV detection, is able to detect CNVs with higher accuracy and resolution, to detect balanced rearrangements such as inversions and translocations, as well as to detect CNVs in regions where probe density of other platforms is low. However, the technique is not feasible for a large number of genomes due to time and budget constraints. Next-generation sequencing (NGS) attempts to combine the benefits of array technology and sequencing. The biggest advantage of NGS over traditional Sanger sequencing is the ability to sequence millions of reads in a single run at a comparatively inexpensive cost (Metzker, 2010). However, with billions of reads generated per individual, there is an increasing need for more bioinformatics support and computers with larger storage and higher computing powers, and for such support to keep pace with the

rapidly changing technologies. Already, there is a great demand for information technology infrastructure and bioinformatics team to analyse the massive amount of data, with speculations that the costs associated with down-handling, storing and analysis of the data could be more than the production of the data.

There is still a need for the development of new statistical/bioinformatics methods and software for the systematic analysis of CNV/SV data. This is the focus of this thesis.

Chapter 2 – BACKGROUND

In this chapter, I will introduce some concepts in CNV/ROH analysis, including definitions and introduction to existing technology, software and algorithms in detection of CNV/ROH. These will facilitate the understanding of subsequent chapters.

2.1 Terminology and nomenclature

Human genetic variations refer to differences in the deoxyribonucleic acid (DNA) sequences among different individuals; they can take many forms, including single-nucleotide polymorphisms (SNPs), indels, copy number variations (CNVs), and other copy-neutral variations such as inversions, translocations and regions of homozygosity (ROHs). These genetic variations span a spectrum of sizes, ranging from 1 base-pair (bp) changes to whole chromosomal changes (e.g. aneuploidy). The occurrences of these genetic variations are attributed to different diverse mechanisms. For example, the predominant mechanisms for CNV formation include non-allelic homologous recombination and non-homologous end joining (Hastings *et al.,* 2009; Conrad *et al.,* 2010). ROHs are thought to be a result of autozygosity or uniparental isodisomy (Gibson *et al.,* 2006).

Table 2.1 summarizes the definitions of variants from single base changes to the sub-microscopic level (larger variants are not discussed). Note that the definitions for the different classes of genetic variants based on size are often unclear at the edges of each class. For example, larger indels may sometimes be termed CNVs even when their sizes are less than 1 kb.

Types of variation	Size	Definition*	Remarks
SNVs, SNPs, single-nucleotide insertions-deletions (indels)	1 bp	SNVs are variations of a single nucleotide (see Figure 1). When the variation is common (usually defined as having a frequency of more than 1%), we call it a SNP (Figure 2.1).	Most SNPs are single nucleotide substitutions, although single nucleotide deletions/insertions may also fall under this category.
Indels, microsatellites, minisatellites, inversions, di-,tri-tetranucleatide repeats, variable number of tandem repeats (VNTRs)	2 to < 1000 bp	Indels are typically defined as insertions or deletions that are smaller than 1 kb and larger than 1 bp.	The size cut off is rather arbitrary; Database of Genomic Variants (DGV) defines indels in the size range of 100 bp to 1 kb.
CNVs, segmental duplications, inversions, translocations	1000 bp to sub-microscopic	CNVs are additions or deletions in the number of copies of a segment of DNA (larger than 1 kb in length) when compared to a reference genome (Figure 2.2).	Some large indels larger than 500 bp may also be termed CNVs. Common CNV larger than 1% population frequency are termed copy number polymorphism (CNP).
ROHs	> 500 bp	ROHs are continuous stretches of the genome (usually more than 500 kb) without heterozygosity in the diploid state.	

Table 2.1: Definition of the different classes of genetic variations, partly adapted from Figure 1 of Scherer *et al.,* 2007. *only selected types of variation are defined.

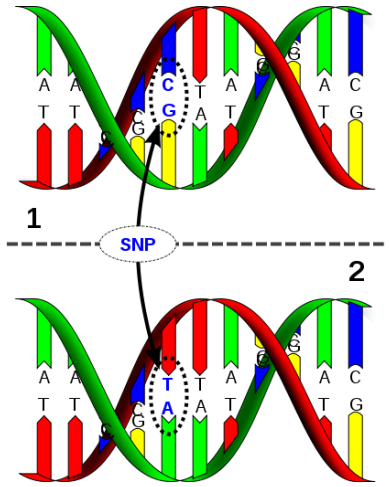


Figure 2.1: C-T single nucleotide variation.Source: http://en.wikipedia.org/wiki/File:Dna-SNP.svg

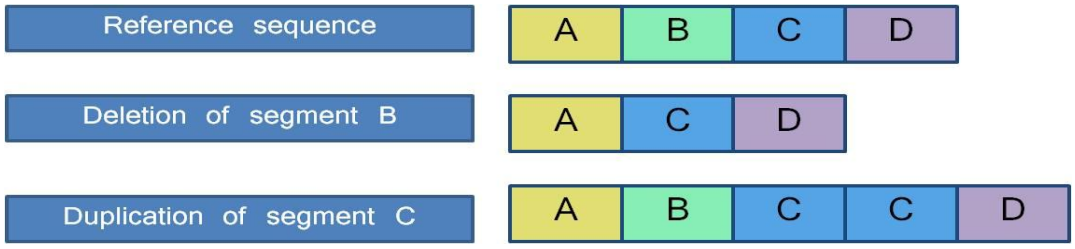


Figure 2.2: Schematic and simplified diagram of a deletion and duplication (adapted from Ku *et al*., 2010).

ROHs are sometimes termed loss of homozygosity (LOH), which includes hemizygous deletions (where there is only one copy of the region). Genotypes of SNPs within hemizygous deletions may be erroneously called as homozygous resulting in a region that may seem to be a ROH based on SNP genotypes alone. Figure 2.3 illustrates the differences in intensity patterns for ROH and one-copy deletion; while both ROH and one-copy deletion have similar B allele frequency (BAF) patterns, the Log R ratio (LRR) for ROH is around zero while it is below zero

for one-copy deletion. In this thesis, ROH always refer to the copy-neutral variant, where the region is in diploid state and all bases within the region are homozygous.

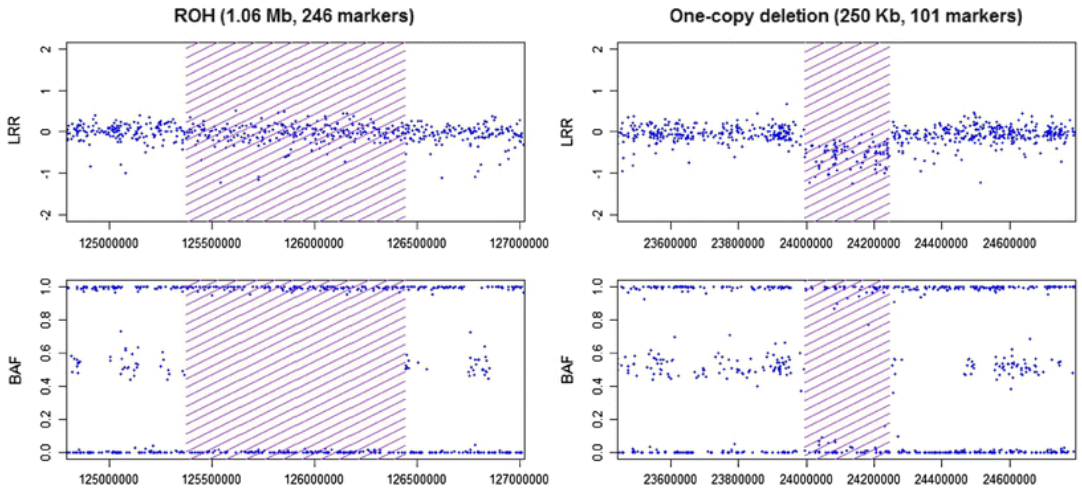


Figure 2.3: (Left panel) ROH signature with LRR around zero and no clusters at BAF of 0.5. (Right panel) One copy deletion signature with decreased LRR and similar pattern of BAF as ROH. The *x*-axis is the genomic probe location and each point represents a probe in the SNP array. (Figure from Ku *et al.*, 2011).

2.2 CNV and ROH detection technologies

In the last decade or so, the most commonly used technologies for CNV detection are whole-genome array comparative genome hybridization (aCGH) and high-density SNP arrays. ROHs are typically detected using high-density SNP arrays. CNVs/ROHs detected using these technologies are unfortunately limited by the density of the probes, as well as the location of the probes. For example, array platforms with more than 1 million probes have a lower detection limit of 10-25 kb in the size of CNV (McCarroll *et al*., 2008). Sanger sequencing provides better resolution and accuracy, but it is not cost/time-effective to use on a genome-wide scale for many individuals. The recent development of next generation sequencing (NGS) platforms that allow massive parallel sequencing have the potential to discover

smaller CNVs that were not previously discovered, detect balanced rearrangements such as inversions and translocations, as well as detect rare CNVs for which SNP arrays have no probes for. The biggest advantage over traditional Sanger sequencing is the ability to produce large amount of sequencing data in a single run.

However, as compared to SNPs, detection of CNVs is more challenging because of its complexity as a multi-base, multi-allelic variant. As a result, different algorithms and methods often give vastly different estimates in the number and breakpoints of CNVs. Currently, in the Database of Genomic Variants (DGV), there are more than 130,000 (merged) CNVs from 37 different studies, encompassing more than 52% of the genome; a likely gross overestimation of the true percentage of the genome encompassed by CNVs. This is because all the different studies use a heterogonous array of technologies, algorithms, filtering parameters, and samples.

2.3 CNV and ROH detection algorithms

Detection of CNVs from aCGH arrays is mostly based on locating change-points in intensity-ratio patterns that would partition each chromosome into several discrete segments. On the other hand, the hidden Markov model (HMM) is particularly popular for detection of CNVs from SNP arrays, where the hidden states provide a natural way of combining information from the total signal intensity (known as log R ratio, LRR) and the relative allele frequency (known as B allele frequency, BAF) values. Briefly, the HMM assumes several possible hidden states such as ‘deletion’, ‘normal’, ‘region of homozygosity’ and ‘duplication’ and analyse the most possible state-transition path, assuming that the copy numbers of nearby SNPs are dependent

(Wang *et al*., 2007). Illustrated in Figure 2.4, a 'normal copy' has three BAF clusters and the LRR is centred around zero; a ROH has LRR centred around zero but only two clusters at both extremes of the BAF.

The output from a CNV detection algorithm provides the following information: (1) Chromosome number (2) Start location (3) End location (4) Copy number. For example, this is a typical output from PennCNV:

chr6:32565228-32593190 numsnp=30 "length=27,963" "state1,cn=0"

It tells us that in Chromosome 6 of this individual, from the position 32565228 to position 32593190, there is a deletion where this individual has zero copies as compared to the reference panel. There are 30 probes in this region in the platform used, and the length of the region is 27,963 bases.

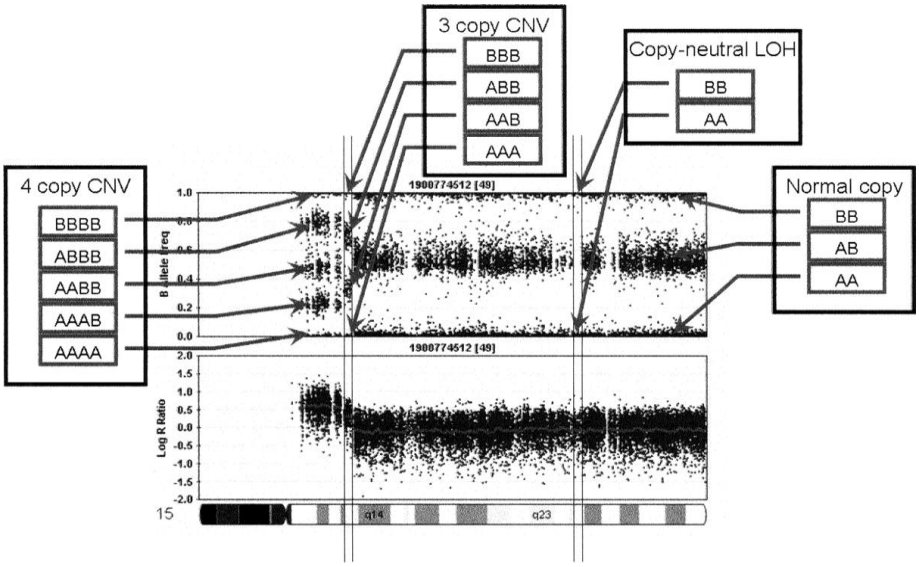


Figure 2.4: Figure from Wang *et al.*, 2007, illustrating the unique patterns in LRR and BAF of the different copy number states. A 'normal copy' has three BAF clusters and the LRR is centred around zero; a ROH has LRR centred around zero but only two clusters at both extremes of the BAF.

2.4 Sequencing technologies

2.4.1 First generation sequencing

First generation sequencing is typically referred to as 'Sanger sequencing', and is introduced by Frederick Sanger in 1977 (Sanger, 1977). It is the main form of sequencing technique used over the last 30 years until the arrival of next-generation sequencers in 2005. Sanger sequencing is able to sequence reads of length ~ 800-1000 bases (Hert *et al.*, 2008; Schloss *et al.*, 2008; Venter *et al*., 2001).

However, Sanger sequencing is laborious and costly; its inability to process more than 96 sequence reads at a time limits its application to large scale genome-wide sequencing efforts for many individuals (Mardis, 2008). For example, it took nearly ten years and three billion dollars to sequence the first human genome in the Human Genome Project (Schadt *et al.*, 2010).

2.4.2 Next-generation sequencing (NGS)

Next-generation sequencing (NGS) or also known as high-throughput sequencing (HTS) is able to simultaneously sequence millions of DNA reads. This ability to produce large amount of sequencing data in a single run at a comparatively inexpensive cost is its biggest advantage over traditional Sanger sequencing (Metzker, 2010). Currently available NGS sequencers in the market include the Roche 454 Genome Sequencer FLX System, Illumina Genome Analyzer, Illumina HiSeq and Applied Biosystems' Supported Oligonucletide Ligation Detection System (SOLiD).

NGS has the potential to discover smaller CNVs that were not previously discovered, to detect balanced rearrangements such as inversions and translocations, as well as to detect CNVs in regions where probe density of other platforms, such as SNP arrays, is low. NGS technologies have facilitated and accelerated the process of identifying genetic variations through whole-genome re-sequencing projects, including the 1000 Genomes Project.

However, there are some technical features of NGS that result in several challenges. Firstly, due to an effect called 'dephasing', there is an increase in noise and sequencing errors as the read length extends, thereby limiting the read lengths of NGS to ~35 – 400 bases (Schadt *et al.,* 2010). The short read lengths in turn complicate alignment and assembly. Secondly, in order to generate a large number of DNA molecules, polymerase chain reaction (PCR) amplification is required. This amplification process biases the frequency in which different portions of the genome are sequenced (Schadt *et al.,* 2010).

2.4.3 CNV detection using NGS

Broadly, there are four complementary methods for CNV detection using NGS data, namely (1) depth of coverage (DOC, also known as read-depth (RD) methods), (2) paired-end mapping (PEM), (3) split-read (SR) and (4) assembly-based (AS) methods (Alkan *et al.,* 2011). Except for the latter, the other three classes of methods require first mapping the sequenced reads to a known reference genome. The different methods are usually complementary to one another as the underlying concepts excel

at detecting certain types of variants, and a large proportion of discovered variants remain unique to a particular approach (Alkan *et al*., 2011).

Some algorithms use a combination of methods for more accurate detection of CNVs. For example, CNVer supplements DOC with PEM information in a unified framework (Medvedev *et al.,* 2010). Genome STRiP combines information from DOC, PEM, SR as well as other features of sequence data at population level (Handsaker *et al.,* 2011). Genome STRiP is one of the highest performing method used in the 1000 Genomes pilot Project, indicating that there is benefit in combining different approaches (Mills *et al*., 2011).

Depth of coverage

DOC methods typically count the number of reads that fall in each pre-specified window of a certain size (Abyzov *et al*., 2011; Yoon *et al*., 2009). The underlying concept of identifying CNVs using DOC is similar is that of using intensity data: a lower than expected DOC /intensity indicates deletion and a higher than expected DOC /intensity indicates duplication (Figure 2.5). The algorithm relies heavily on the assumption that the sequencing process is uniform, i.e., the number of reads mapping to a region is proportional to the number of copies. However, certain biases such as GC-content and mappability cause this assumption to be unrealistic; regions of the genome may be over or under-sampled regardless of the copy number of the region, often resulting in spurious signals. DOC algorithms usually detect large CNVs and are unable to detect copy neutral events such as inversions and translocations. Single-end or paired-end data may be used for this analysis.

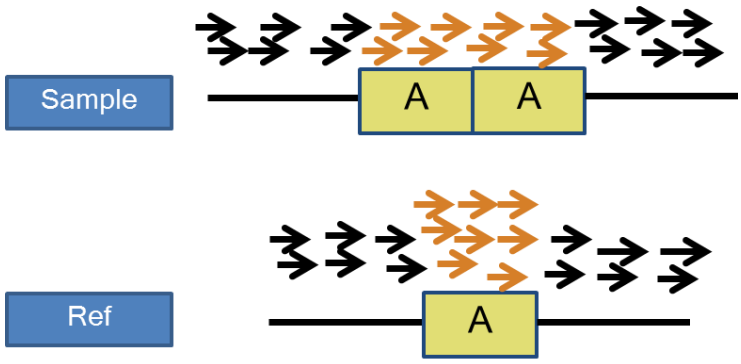


Figure 2.5: Schematic diagram illustrating the concept of depth of coverage method for CNV detection. If the sample has an additional copy relative to the reference genome, when the reads are mapped to the reference, we would observe an increase in depth of coverage in the region.

Paired-end mapping

PEM methods require the reads to be paired (Chen *et al.*, 2009). The concept is that the fragments of DNA from which the reads are to be sequenced have a fragment length (or also known as insert size) of a certain distribution, and a longer than expected fragment length indicates a deletion in the studied genome compared to a reference while a shorter than expected fragment length indicates an insertion. Based on the patterns from which the paired reads are mapped to the reference, read pair analysis can also detect inversions and translocations. The size of CNVs detected using PEM is limited by the insert size and as a result, PEM often detects smaller CNVs. For example, PEM does not allow the discovery of insertions larger than the insert size (Dalca *et al.*, 2010).

Split-read

SR methods uses paired reads as well. They focus on pairs of reads where one read is mapped to the reference while the other read failed to be aligned (Ye *et al.*, 2009).

The idea is that where the location of the unmapped read may span the breakpoint of the CNV. SR analysis has the advantage of being able to pinpoint the location of the breakpoints.

Assembly-based

AS methods, on the other hand, do not align the reads to a known reference but construct the genome piece-by-piece, which is known as *de novo* sequencing. Some AS methods use the reference genome as a guide to resolve repeats. This is known as *comparative assembly* (Pop *et al.,* 2004). AS methods can discover new non-reference sequence insertions. AS methods works best for small genomes such as bacterial genomes and are less widely used in NGS sequencing of humans because the short reads from NGS makes assembly in repeat regions difficult (Ye *et al*., 2009). Even though assembly algorithms continue to improve, due to the short read lengths, de novo sequencing using NGS are still not capable of achieving similar quality as that using Sanger sequencing (Schadt *et al*., 2010).

2.5 Repetitive DNA

Repetitive DNA refers to sequences that are highly similar or identical to sequences in other parts of the genome. They are abundant in the human genome and covers almost 50% of the human genome (Treangen *et al.,* 2012). Table 2.2 summarises repeat type, number, percentage of genome covered and approximate length of each repeat class. The repeat type is broadly characterized into tandem or interspersed repeats where the former refer to repeats that are adjacent to each other while the latter refers to repeats that are separated by hundreds, thousands or millions of bases.

In next generation sequencing, reads from repetitive regions may map equally well to several locations in the reference genome. Due to the ambiguity in the alignment step, these reads often cause problems in SNP and SV detection. Reads that can be mapped equally well to more than one location are termed *multi-reads*.

Repeat class	Repeat type	Number	% genome	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426, 918	3%	2 -100
SINEs	Interspersed	1, 797, 575	15%	100 - 300
DNA transposon	Interspersed	463, 776	3%	200 - 2000
LTR retrotransposon	Interspersed	718, 125	9%	200 - 5000
LINEs	Interspersed	1, 506, 845	21%	500 - 8000
rDNA	Tandem	698	0.01%	2000 - 43000
Segmental duplications and other classes	Tandem or interspersed	2, 270	0.20%	1000 - 100000

Table 2.2: This table summarises for each repeat class, the repeat type (tandem or interspersed), number in the hg19 human genome, percentage of the hg19 human genome covered, and approximate lower and upper bounds for the lengths of the repeat. (Table adapted from Treangen *et al.,* 2012). Short interspersed nuclear elements (SINEs), Long terminal repeat (LTR), Long interspersed nuclear elements (LINEs), ribosomal DNA (rDNA).

2.6 Copy number variation region (CNVR)

CNVR or also known as CNV loci or common CNV or recurrent CNV are CNVs that occur in the same/similar location across several individuals. Most CNV detection algorithms identify CNVs individual-by-individual, but common CNVs are known to exist among different individuals. However, the identification of the individual-specific CNVs is not precise, especially in terms of the breakpoints. This poses a challenge when we want to summarize the population characteristics or perform association studies, because it is unclear if CNV1 from individual 1 describes biologically the same event as CNV2 from individual 2 if their breakpoints do not match exactly.

2.7 Hardy Weinberg Equilibrium of CNVR

Suppose a bi-allelic SNP has allele frequencies p and q (where $p+q = 1$) for alleles a and b respectively, regardless of gender. Assuming random mating, in the next generation, the frequencies of the aa, ab and bb genotypes are p^2, $2pq$ and q^2 respectively. The allele frequencies of a and b have not changed and remain p and q, such that in the following generation, the genotype frequencies will again be p^2, $2pq$ and q^2, and so forth. This is known as Hardy Weinberg Equilibrium (HWE); i.e., that the frequency of alleles and genotypes remain constant from generation to generation in a large population assuming random mating. The Pearson's chi-squared test is typically used to test for departure from these expected frequencies, indicating violation of HWE.

Since it has been observed that the majority of common CNV regions are inherited (Locke *et al.*, 2006), we expect, for a population of normal, healthy individuals, the integer copy numbers for the majority of CNVRs to be in HWE. This is supported by McCarroll *et al*., (2008)'s study that found that 98% of common bi-allelic CNVRs do not violate HWE. McCarroll *et al*., (2008) also found that about 90% of common CNVs are bi-allelic.

In principal, HWE applies to both bi-allelic CNVRs and multi-allelic CNVRs. Bi-allelic CNVs are those with only two alleles, forming three possible copy numbers. For example, CNVs with copy numbers 0, 1, 2 or 2, 3, 4 are considered bi-allelic. Multi-allelic CNVs are those with more than two alleles, for example, with alleles '0', '1' and '2', the possible copy numbers are 0, 1, 2, 3 and 4. Testing HWE for bi-allelic

CNVs is straightforward and similar to the test for SNPs. However, for multi-allelic CNVRs, HWE test cannot be performed directly on the unphased copy-number because there is an issue with different combinations of alleles producing the same copy-number. For example, with alleles '0', '1' and '2', the copy number 2 can have genotype (1, 1) or (0, 2).

2.8 GWAS of CNVs

Genome-wide association studies (GWAS) using SNPs have been widely performed over the last couple of years, resulting in over 1400 published associations (at $p \leq 5x10^{-8}$) for 237 traits (from the National Human Genome Research Institute: http://www.genome.gov/26525384). This is in part due to greater accuracy and completeness with which SNPs, as compared to CNVs, can be assayed.

Earlier studies on CNV discovery have paved the way for subsequent association studies of CNVs. For example, the Wellcome Trust Case Control Consortium (WTCCC) performed a large scale GWAS study of CNVs in 16000 cases of eight common diseases using a customized aCGH that was designed based on previously identified CNVs (Wellcome Trust Case Control Consortium, 2010). The WTCCC study found several CNV loci to be associated with Crohn's disease, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. However, these loci have been previously identified through SNP based GWAS, reflecting the observation that common CNVs are well tagged by SNPs.

2.9 Linkage disequilibrium

The non-random association of alleles at two or more loci in the genome is known as linkage disequilibrium (LD), i.e. that the occurrences of some combinations of alleles at two or more loci are more or less frequent than expected based on their individual allele frequencies. For example, suppose allele A_1 at SNP A and allele B_1 at SNP B have frequencies p_1 and q_1 respectively. If the two SNPs are independent, then we expect to see the A_1B_1 haplotype at a frequency of p_1q_1; any departure from this expected frequency means that the two SNPs are in LD. Most commonly used statistics to quantify the extent of LD between two loci are the r^2 and D' statistics (Lewontin *et al*., 1960). Both statistics are based on the extent of departure of the observed haplotype frequency from the expected. Let x_{11} be the observed AB haplotype frequency. Then, $D = x_{11} - p_1q_1$. Now, let the other two alleles of SNP A and SNP B have frequencies p_2 and q_2 respectively.

$$r^2 = \frac{D^2}{p_1p_2q_1q_2} \text{ and } D' = \frac{D}{D_{max}} \text{ where } D_{max} = \begin{cases} \min(p_1q_1, p_2q_2) & \text{when D} < 0 \\ \min(p_1q_2, p_2q_1) & \text{when D} > 0 \end{cases}$$

Both measures have a minimum value of 0, which indicates independence between the two loci, and maximum value of 1, which indicates complete dependence between the two loci.

2.10 Quantification of positive selection

Positive selection is the phenomenon where certain variants rise to a frequency at a faster rate than would be expected, i.e., the favouring of variants that increase survival and reproduction. Under neutral evolution, new variants need a long time to reach

high frequency, resulting in common variants usually having short range LD because recombination would have occurred to disrupt the haplotypes (Sabeti *et al.,* 2002). Hence, one 'clue' or signature that provides evidence of positive selection is an unusually long and common haplotype which indicates an allele which rose to high frequency rapidly before recombination occurs (Bersaglieri *et al.*, 2004).

One statistic used to quantify positive selection is the integrated haplotype score (iHS) (Voight *et al.*, 2006). Briefly, this score measures how unusual the haplotypes around a core SNP are, relative to the rest of the genome. The iHS first utilizes the extended haplotype homozygosity (EHH) statistic (Sabeti *et al.,* 2002); the EHH measures the decay of haplotype identity as a function of distance. For each SNP, haplotype homozygosity starts at 1 and decays to zero with increasing distance. Alleles under selection tend to have high haplotype homozygosity that extends much further, resulting in a large area under the EHH curve. The iHS is a standardized measure of the integrated EHH. Clusters of SNPs with large positive or large negative iHS are evidence of position selection in the region.

Chapter 3 – AIMS

Overall, the general aim of this thesis is to use and develop statistical and bioinformatics methods to improve detection and analyses of structural variants. The thesis is divided into four studies as follows:

I. We develop a method and accompanying software to identify common CNV regions in multiple individuals. The identified common regions can be used for downstream analyses such as group comparisons in association studies.

II. We develop a method and software to identify CNVs by using data from multiple platforms simultaneously. We also propose an objective criterion for discrete segmentation required for downstream analyses. For each identified segment, the software reports a p-value to indicate the likelihood of the segment being a true CNV.

III. We investigate the population characteristics of ROHs in three Singapore populations (Chinese, Malays and Indians), and access the relationship between the occurrence of ROHs and haplotype frequency, regional LD and positive selection.

IV. We highlight problems and issues encountered when analysing NGS data for CNVs, in particular, those pertaining to DOC methods. We use real data from the 1000 Genomes Project to highlight and investigate challenges associated with (1) GC-content, (2) quality score of reads, and (3) identifying CNVs in repeated regions.

Chapter 4 - PAPER SUMMARIES

4.1 Study I: Identification of recurrent regions of copy-number variation across multiple individuals.

4.1.1 Motivation

Most algorithms for CNV-detection detect CNVs sample-by-sample with individual specific breakpoints. However, common CNV regions (CNVRs) are likely to occur at the same genomic locations across multiple individuals.

4.1.2 Methods overview

The main novelty of our algorithm is that we exploited the region specific confidence score statistic provided by commonly used segmentation programs, PennCNV and QuantiSNP. This statistic indicates how likely the detected CNV for a particular individual is true. By not incorporating the use of individual specific confidence scores, it means that all regions contribute equally to the statistic used to identify the common regions, but some regions are more likely to be true positives than others. Our method utilizes both the confidence score statistic, as well as the frequency of occurrence, to identify CNVRs. Intuitively, we have less confidence in a CNV that occurs in one individual than one that occurs in many individuals. However, a single occurrence of CNV might still be a true discovery if it is associated with a high confidence score, i.e., it is based on a strong signal. Since individual CNVs span different probes, the number of individual regions that overlap each probe varies. However, common CNV regions tend to occur at almost the same genomic locations

across multiple individuals. Hence, we expect the common regions to be identified by consecutive probes where a 'significant' number of individuals have an overlapping CNV. Furthermore, we also expect the confidence score of the individual regions to be relatively high.

Method 1: Cumulative Overlap Using Very Reliable Regions (COVER)

To calculate the COVER statistic for a probe, we sum the number of high-confidence individual regions that overlap that probe. The common region is then defined as consecutive probes for which the COVER statistic is greater than or equal to a specified threshold, *u*. Users provide two parameters here: the confidence score threshold, *c*, to determine high-confidence regions and *u*, the threshold for the COVER statistic.

Method 2: Cumulative Composite Confidence Scores (COMPOSITE)

In COVER, we may miss regions that are detected with lower confidence scores but nonetheless detected consistently across a large number of individuals. For the COMPOSITE statistic, we sum all individual regions that overlap the probe, weighted by their confidence score.

Method 3: Clustering of Individual CNV regions within a Common Region

The CLUSTER method uses a clustering algorithm that further refines the regions identified by either method 1 or method 2. This method is motivated by the observation of a complex mixture of sub-regions within a CNVR identified by COVER/COMPOSITE (Figure 4.1).

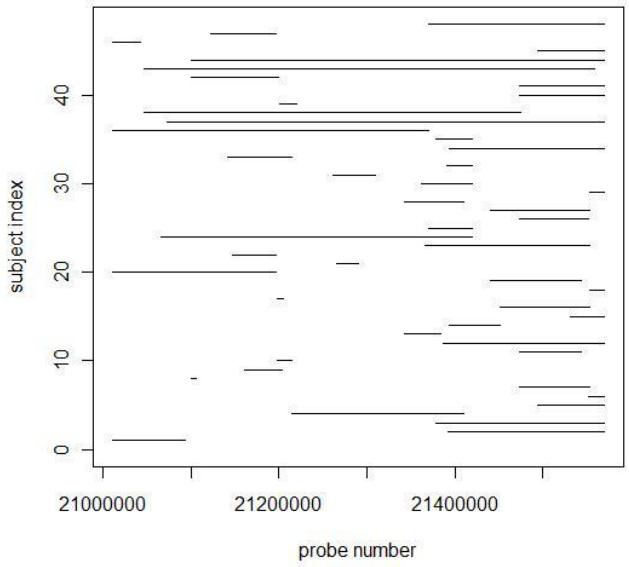


Figure 4.1: An example of a CNVR identified by COVER. We observe that despite being identified as a common region, the individual regions still portray a mixture phenomenon of several distinct sub-regions (from Teo *et al*., 2010).

4.1.3 Results

Comparison with sequenced regions

To assess the performance of our methods, we use 112 HapMap samples and vary the threshold parameters in our methods. For each threshold, we calculate discordance rates with sequencing-based results (Kidd *et al.,* 2008) and rates of departure from HWE. The discordance rates as well as the rates of departure from HWE decrease when we select CNVs with higher confidence scores, showing the importance of further processing of the CNVs (for COVER results, see Figure 4.2). Similar results were observed for COMPOSITE method (Figure not shown). Concordance rates improve after refinement with CLUSTER.

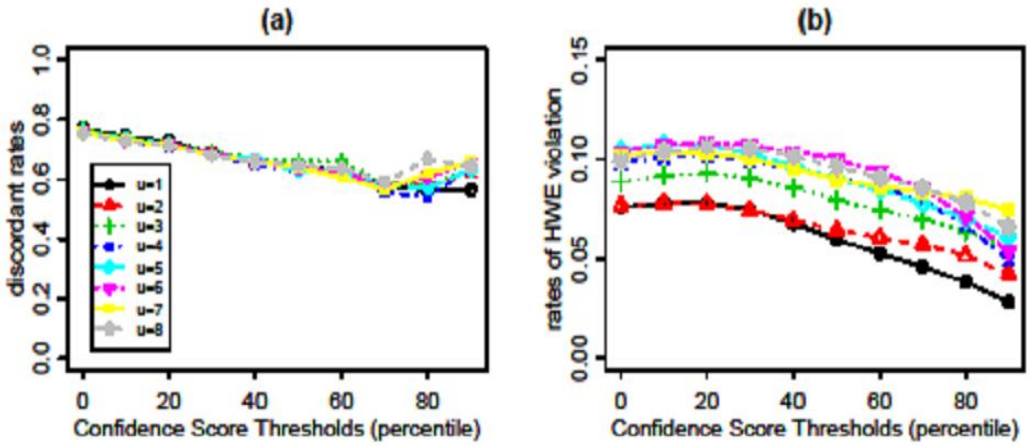


Figure 4.2: (a) Discordance rates for COVER method decreases as the confidence score thresholds increase. (b) Rates of departure from HWE decreases as the confidence score thresholds increase (from Teo *et al*., 2010).

Comparison to other algorithms

We compare our methods to two previously published methods, STAC (Diskin *et al*., 2006) and GISTIC (Beroukhim *et al.,* 2007). We find that our methods are better at identifying low-frequency but high-confidence CNV regions.

Implementation

The methods are implemented in an R package, *cnvpack*. The main input is a list of detected individual CNV regions with the following information: Sample name, chromosome number, detected integer copy number, start and end genomic locations and a confidence score. The package can be downloaded from http://www.meb.ki.se/~yudpaw.

4.2 Study II: Multi-platform segmentation for joint detection of copy number variants.

4.2.1 Motivation

At the time this research was carried out, SNP genotyping platforms from major commercial companies, such as Illumina and Affymetrix, were rapidly evolving, and it is not uncommon for research groups to have data from different platforms for the same individuals. For CNV detection, marker density is one important factor. Different platforms have different sets of marker panels and combining data from multiple platforms would undoubtedly give higher marker density. It has the potential to yield more precise and accurate detection of CNVs and its breakpoints. However, combining such data is not straightforward as different platforms show different degrees of attenuation of the true copy-number, noise characteristics and marker panels (Zhang *et al.*, 2010). There is still a relative lack of formal procedures for combining information from different platforms for copy-number calling. Most studies with data from multiple platforms interrogating the same samples usually process the data independently for each platform, after which the identified segments are combined in an ad-hoc manner. This approach does not fully utilize information from the different platforms, and when the segmented results from the different platforms differ, it is difficult for researchers to come to a consensus in a statistically rigorous manner.

In this study, we develop a new method for identifying CNVs by using data from multiple platforms simultaneously. As we are often interested in discrete segments of

CNVs for downstream analyses, we also develop an objective method to obtain discrete segments, and provide a p-value associated with each segment; the p-value would indicate how likely the segment is a CNV, and can be used to filter false positives.

4.2.2 Methods overview

The method, multi-platform smooth segmentation (MPSS) is an extension of Huang *et al*. (2007)'s single-platform *smoothseg* algorithm which is based on the Cauchy random-effect model that allows jumps in the underlying copy-number patterns. MPSS uses normalized $\log_2$-intensity ratios from two or more platforms and estimates the underlying copy number pattern for an individual. For each individual, we denote $\{x_1, \ldots, x_n\}$ as the union of the probe locations from the different platforms, with $x_1 < x_2 < \ldots < x_n$. Denote $\{y_{x1j}, \ldots, y_{xnj}\}$ as the set of $\log_2$-intensity ratios from platform j. We write our model as

$$y_{xij} = f(x_{ij}) + e_{xij} ,$$

where f is a random effects parameter that is common to all platforms, meaning that each platform is assumed to measure the same underlying copy-number pattern; as such, background normalization is recommended so that data from the different platforms become comparable. The error term e_{xij} is platform-specific to take into account different noise characteristics of the different platforms. The platform specific error structure was chosen to be *t*-distributed to incorporate a heavy tailed structure that can deal with outliers in the observations. The smoothness of f can be

expressed by assuming that the scaled second order differences $a_i^* \equiv \frac{\Delta^2 f_t}{(\Delta x_t)^2}$ are independent and identically distributed with some distribution. We specify a_i^* to follow the Cauchy distribution to allow for jumps in the segments. To estimate the random-effects parameter *f*, we derive an iterative weighted least squares algorithm by maximizing the likelihood of the Cauchy random-effects model.

4.2.3 Results

We compare *MPSS* against the single-platform *smoothseg* algorithm, an existing multiplatform method, called *MPCBS* (Zhang *et al*., 2010), and its associated single platform method, *CBS* (Olshen *et al*., 2004). We use nine HapMap samples, which were previously genotyped by both the Illumina 1M and Affymetrix 6.0 SNP arrays by our collaborators at the Genome Institute of Singapore, Agency for Science, Technology and Research. For the same samples, we have the integer copy-numbers from Conrad *et al.* (2010)'s study, which we use as a reference list.

When signals from the different platforms are consistent, we get increased power to detect the CNVs when we combine information from different platforms, especially in areas where a single platform has low density of probes (Figure 4.3a) or complete lack of probes (Figure 4.3b). To compare against other methods, we perform individual-specific comparisons with the reference list and report the number of overlapping bases as a proportion of the total length of CNVs identified by the method and as a proportion of the total length of CNVs in the reference list. In Figure 4.4, we show that MPSS CNVs have greater amount of overlap with the reference, indicating better performance.

Implementation

The algorithm is implemented in an R package *MPSS* that can be freely downloaded from http://www.meb.ki.se/~yudpaw. The main inputs are vectors of genomic positions, chromosome numbers and $\log_2$-intensity ratios from each platform.

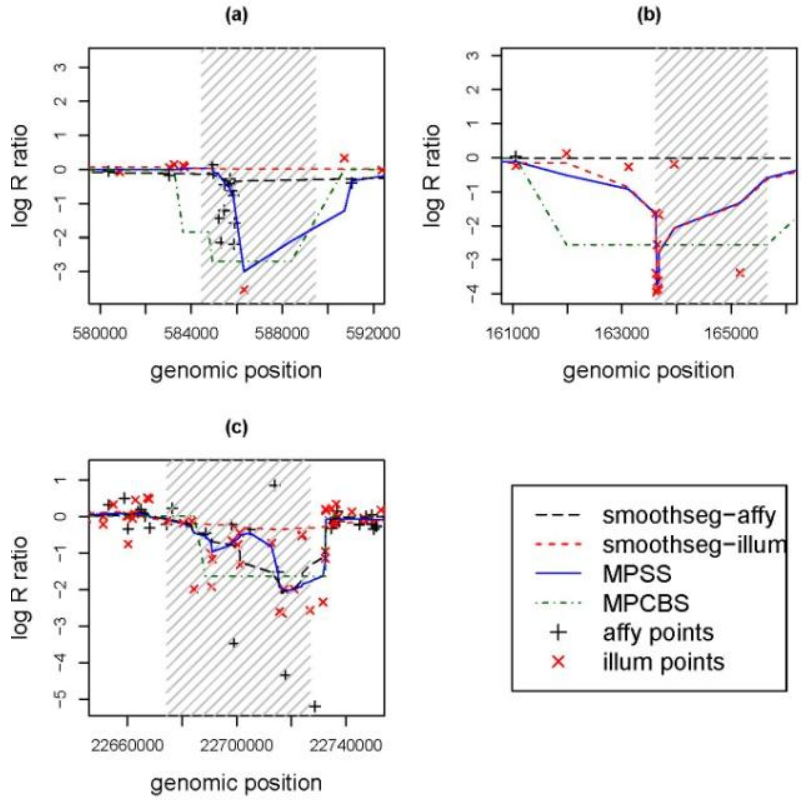


Figure 4.3: Examples of segments detected by the multiplatform methods. (a) A deletion in Chromosome 8. Single platform smoothseg on Illumina platform was unable to identify the deletion due to lack of probes in the region. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to insufficient signal. (b) A deletion in Chromosome 16. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to complete lack of probes in the region. (c) A deletion in Chromosome 22 (from Teo *et al.,* 2011).

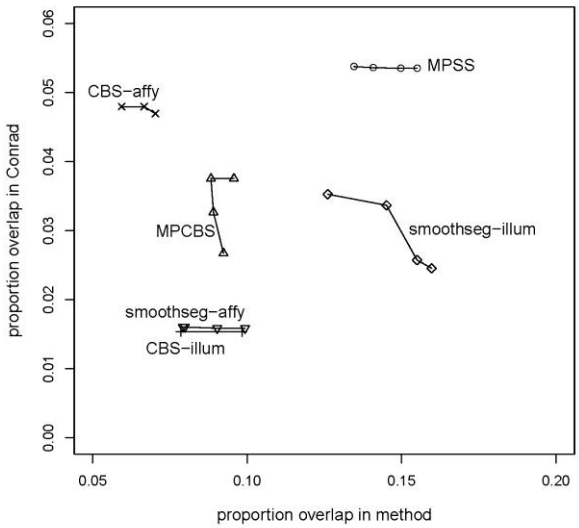


Figure 4.4: The number of overlapping bases as a proportion of Conrad's CNVs and as a proportion of each method's CNVs; the different points for each method correspond to the different thresholds. A higher proportion of overlap indicates better performance (from Teo *et al.,* 2011).

4.3 Study III: Regions of homozygosity (ROHs) in three Southeast Asian populations

4.3.1 Motivation

The genomes of outbred populations were first shown in 2006 to contain an abundance of long stretches > 500kb without heterozygosity (Gibson *et al.,* 2006; Li *et al.,* 2006). Since then, there have been several studies that investigate the population characteristics of ROHs in healthy individuals (McQuillan *et al.,* 2008; Nothnagel *et al.,* 2010; O'Dushlaine *et al.,* 2010), and also several studies that perform association analyses to identify ROHs that are associated with complex diseases (Yang *et al.,* 2010; Lencz *et al.,* 2007; Nalls *et al.,* 2009). However, the majority of these studies are conducted on European populations, and there is a lack of knowledge of ROHs in Asian populations. Thus, the first aim is to characterize ROHs in the three main Singapore populations, namely the Chinese, Malays and Indians.

Furthermore, it was observed that the location of ROHs is markedly non-random, where unrelated individuals may share similar region boundaries. Some loci are caused by a single common haplotypes, whereas others are a consequence of several common haplotypes that could be markedly disparate (Curtis *et al.,* 2008). The second aim of this study is to investigate the relationship between the occurrence of ROHs and haplotype frequency, linkage disequilibrium (LD) and positive selection.

4.3.2 Samples

The genomic DNA samples used in this study were part of the Singapore Genome Variation Project[1], whose aim was to characterize the extent of common genetic polymorphisms and the haplotypes in each of the three ethnic groups in Singapore (Teo *et al.,* 2009). Peripheral blood DNA was extracted from a total of 292 individuals and genotyped using the Illumina Human 1M Beadchip and the Affymetrix Genome-wide Human SNP Array 6.0.

4.3.3 Results

We identified an average of 207, 179 and 126 ROHs per individual for Chinese, Malays and Indians respectively. Indians have lower numbers as well as lower total length of ROHs as compared to Chinese and Malays. About 83% of the ROHs are within the 500 kb to 1 Mb size range while 17% of them are greater than 1 Mb.

Using the individual regions to form common regions (using the software from Study I), we obtain 1256 common ROH loci in the three populations. We study the relationship of the common ROHs with haplotype frequency, LD and positive selection. For each locus, we test for differences among the 3 populations in terms of ROH frequencies and haplotype frequencies, and 47 loci (<4%) differ significantly in frequencies while 899 loci (69%) differ significantly in haplotype frequencies among the populations. One interesting example is a 700 kb region in Chromosome 16 that overlaps with the Vitamin K epoxide reductase complex subunit 1 (VKORC1) gene,

[1] approved by the National University of Singapore – Institutional review Board (Reference Code: 07 – 199E)

where genetic polymorphisms within this gene has been found to correlate with differences in warfarin dosage and response (Aquilante *et al*., 2006; Harrington *et al*., 2005). In the Singapore populations, the Indians were observed to display warfarin resistance, thus requiring a higher dose as compared to the Chinese and Malays (Zhu *et al.,* 2007). The ROH frequencies of this region are 21%, 13% and 20% for the Chinese, Malays and Indians respectively (no significant difference in frequencies). However, the haplotypes frequencies of this region among the three populations differ drastically (Table 4.1), especially between the Indians and the other two populations.

	Haplotype A	Haplotype B
Chinese	0.31	0.0052
Malay	0.28	0.045
Indian	0.0060	0.34

Table 4.1: Haplotype frequencies of three populations in an ROH that overlaps VKORC1 gene (from Teo *et al.,* 2012).

With regards to haplotype frequency and regional LD, we find that the frequency of an ROH is positively associated with the total frequency of the top three haplotypes as well as with regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis (PCA), demonstrating that ROHs contain information on population structure and the demographic history of a population.

4.4 Study IV: Statistical challenges associated with detecting CNVs using next-generation sequencing (NGS) technology.

4.4.1 Motivation

Whole genome re-sequencing for the identification of CNVs has gained popularity with the recent development of NGS platforms that allow massive parallel sequencing. These techniques have the potential to discover smaller CNVs that were not previously discovered and detect balanced rearrangements such as inversions and translocations. However, analysing NGS data for CNVs is a new and challenging field, with no standard protocols or quality control measures. Also, due to the complexity of the genome and the short read lengths from NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs, no matter which method or algorithm is used.

4.4.2 Results

We describe and discuss areas of potential biases in CNV detection using NGS data, focusing on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulties in identifying duplications. To gain insights to some of the issues discussed, we download real data from the 1000 Genomes Project and analyse its depth of coverage (DOC) data. We show examples of how reads in repeated regions can affect CNV detection, demonstrate current GC correction algorithms, investigate sensitivity of DOC algorithm before and after quality-control of reads and discuss reasons for which duplications are harder to detect than deletions.

Chapter 5 - DISCUSSION

5.1 What makes a good CNV detection method?

The quality of a CNV detection method (including the technology and algorithm) can be broadly attributed to three aspects: (1) location (2) breakpoints (3) genotype. The location and breakpoints of a CNV are closely related, where the breakpoints are given by the start and end positions of a CNV, and the location is the entire region that spans from the start position to the end position. Most studies use the location and breakpoints to determine sensitivity and specificity of a method. However, with SNP/aCGH arrays, the start and end positions are technically not the true start/end positions of a CNV, but rather the start and end probes of the array that was used. Hence, breakpoint precision is highly affected by the resolution of the array. An array with denser probes at and near the location of the CNV will be able to detect the start/end of the CNV with higher precision.

Another less-frequently used criteria for evaluating CNV detection methods is the ability to discern the actual copy number of the region, for example 0 copy versus 1 copy for deletions and 3 or more copies for duplications. This is also known as 'genotyping' of the CNV. Many algorithms use a clustering procedure, assuming that most individuals have normal '2 copies'.

5.2 Concordances among CNV detection methods

From experience of several peer reviews we got during our submission of the manuscripts, many reviewers are often concerned about the low concordance between

the CNVs generated by our methods as compared to the reference list we use. However, this low concordance is often not a very good indicator of bad algorithm performance per se, but rather a more general problem in CNV detection. For example, in McCarroll *et al.* (2008)'s study, they employed a set of very strict criteria on duplicate experiments in SNP arrays to define common CNV regions in eight HapMap samples. Despite that, (on average) 76% of the regions do not overlap with the list of regions found using sequencing. Even when applied to the same raw data, Pinto *et al.* (2011) found that different analytic tools typically yield CNV calls with <50% concordance. The low concordance can be attributed to several factors such as (1) lack of a true gold standard, (2) noisy data resulting in many false identifications and (3) imprecision of the breakpoints identified.

Indeed, the first step of determining the sensitivity of a method is to obtain a 'true positive' dataset. Hence, the first problem with CNV analysis: we do not know the 'true positives'! The closest bet is to use published results from studies that are well-validated as a reference panel, and that is often only possible if you have the same samples as that in the reference panel. HapMap samples are commonly used in methodology research, usually for two main reasons: the raw data are readily available and there are several studies which have characterized the CNV profiles for these individuals and often used as the 'gold standard' (Kidd *et al*., 2008; McCarroll *et al*., 2008; Conrad *et al.*, 2010). When this is not possible, simulation is another way to estimate the sensitivity of the method.

After we have chosen our 'gold standard' dataset, the second difficulty in accessing sensitivity is in answering the question "Is CNV1 and CNV2 the same variant?" In

Figure 5.1a, when the breakpoints of the two variants match perfectly, there is no doubt in calling them the same variant. In Figure 5.1b, the breakpoints are different but the two variant have a good amount of overlap and are of roughly the same length. What about in Figure 5.1c where one breakpoint coincides but the length of the variant differs by a lot? Some studies use a relaxed criterion of calling two variants the same as long as there is a single base overlap, while other studies may be as stringent as requiring at least an 80% reciprocal overlap. A 50% reciprocal overlap seems to be adopted by the majority of studies in recent years. To avoid the need to choose this arbitral percentage, some studies define sensitivity as the proportion of bases that overlap.

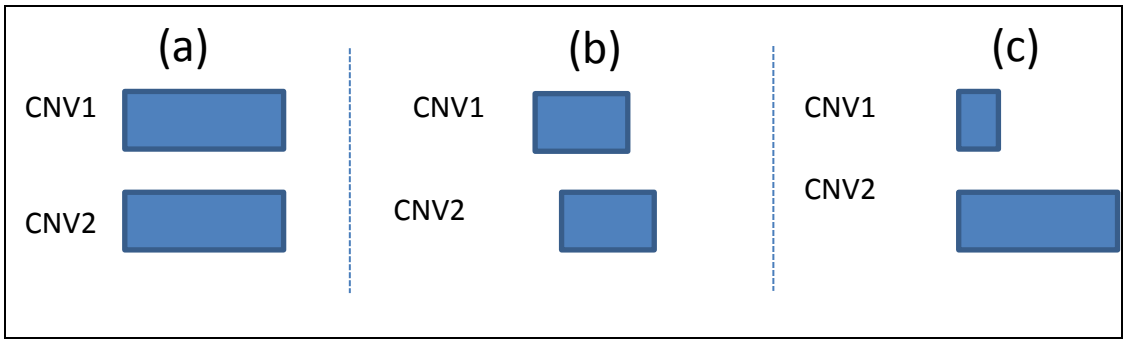


Figure 5.1: Diagram illustrating the non-triviality of determining if two CNVs are the 'same' variant. In (a), CNV1 and CNV2 overlap completely. In this case, we are confident that the two CNVs are the same. In (b), the start and end positions of CNV1 and CNV2 differs, but there is substantial overlap between the two. In (c), CNV1 is completely within the range of CNV2 but the two CNVs differ vastly in lengths. In most research papers, scientists are comfortable with using a 50% reciprocal overlap to determine if two CNVs are concordant.

5.3 Problems caused by repetitive DNA

Repetitive DNA poses challenges in CNV detection regardless whether SNP arrays or sequencing methods are used. For SNP arrays, the density of SNP probes in segmental duplicated regions is sparse due to technical difficulties in assay design and

implementation (Winchester *et al.*, 2009) resulting in a bias against detecting CNVs in segmental duplicated regions using SNP arrays.

For sequencing, reads that fall in repetitive DNA cause problems in alignment and assembly algorithms (Treangen *et al.*, 2012). This problem is exacerbated in NGS (as compared to Sanger sequencing) because the sequenced reads from NGS are relatively short (35-150bp). Furthermore, mutations or sequencing errors in one or two locations may also cause reads to be mapped wrongly (Li *et al.*, 2008). In the 1000 Genomes trios Project, about 20% of the reference genome was considered inaccessible (defined as regions with many ambiguously placed reads or unexpectedly high or low numbers of aligned reads). The resulting low sensitivity in detecting CNVs in repeated/segmental-duplicated regions is a serious problem, because there is an observed enrichment of CNVs in segmental duplicated regions and many breakpoints lie in duplicated regions (Medvedev *et al.*, 2009).

For assembly-based methods, repeat regions create challenges because if the read length is shorter than the repeat region, it is not straightforward to decipher the original sequence since overlap between the reads or contigs will be ambiguous (Knudsen *et al.*, 2010). For other methods that require mapping to a reference, there are different alignment strategies for dealing with multi-reads, such as (1) discarding the reads, (2) choosing a position at random out of all equally good match positions, and (3) reporting all possible positions. In Study IV, we have shown why these strategies are inadequate for dealing with multi-reads.

Recently, there are several algorithms that claim to be able to resolve specific types of CNVs in repeat regions. For example, He *et al*. (2011) developed an algorithm for tandem copy number variation reconstruction in repeat-rich regions, which considers all locations of possible mappings and uses information on read-pair and DOC. Alkan *et al.* (2009) developed a new alignment method, mrFAST. The aligner maps short sequence reads to a repeat-masked reference genome, meaning that all loci with known high-copy common repeats were first masked before alignment, and reports all mapping locations for multi-reads. It also keeps track of mutation in multi-reads. This method has been shown to be able to predict absolute copy number and multicopy differences. Sudmant *et al*. (2010) also uses a similar approach to identify and genotype CNVs within segmental duplications. However, these approaches seem to work only for deeply sequenced data (>20X), and more has to be done to extend these methods for lower coverage data (Chiang *et al*, 2009).

Longer read lengths from third generation sequencing may partially solve the problems with repeats, but even with a read length of 1kb, there still remains about 1.5% of the human genome sequence that is non-unique (Schatz *et al.*, 2010).

5.4 A peek into third generation sequencing (TGS)

Third generation sequencing (TGS) or also known as single molecule sequencing (SMS) promises to improve sequencing rates, throughput and read lengths as compared to NGS. Since it does not require repeated stepwise 'washing and scanning' procedures like in NGS, TGS may increase the sequencing cycle by four orders of magnitude (Eid *et al*., 2009). The first commercially available SMS instrument is the

HeliScope Single Molecular Sequencer by Helicos Biosciences; however, the read lengths are still short at ~32 bases long (Schadt *et al*., 2010). Since PCR amplification is not required in TGS, bias observed in NGS in depth of coverage due to PCR may be resolved. The longer read lengths of TGS will also improve challenges caused by the short read lengths of NGS. Time will reveal if TGS can fulfil its promises for advancement over NGS.

CHAPTER 6 - CONCLUSIONS

- Copy number variations, ROHs and other structural variations are an important source of variation in the human genome, and have been associated with many complex diseases.
- Due to the multi-base and multi-allelic nature of these variants, detecting them with high sensitivity and specificity is still a challenge. Hence, new statistical methods and user-friendly bioinformatics tools are needed for the analyses of these variants.
- In Study I, we develop a method that allows users to detect common CNV regions.
- In Study II, we develop a method that allows users to detect CNVs using information from multiple platforms simultaneously.
- There is a lack of studies investigating regions of homozygosity in Asian populations. There is also a lack of understanding of the relationships between ROHs and haplotype frequency, linkage disequilibrium and positive selection. These are addressed in Study III.
- Next-generation sequencing has the potential to detect CNVs beyond the resolution of SNP arrays and aCGH, as well as detect copy neutral SVs such as inversions and translocations.
- Analytical methods and algorithms for CNV detection using NGS are not yet mature and there are still many challenges. In Study IV, we describe and discuss challenges faced in CNV detection using NGS data.

Chapter 7 – FUTURE DIRECTIONS AND PERSPECTIVES

The field of genetics and genomics has progressed a long way since the first human genome was sequenced in 2000. By now, there are thousands of genes and loci discovered that are associated with simple and complex human diseases, and many of the discoveries were made via GWAS of SNPs. SVs, on the other hand, were much less considered in association studies, particularly attributed to technical difficulties in characterizing SVs with high resolution. Recent development of high-throughput sequencing presents new opportunities for identifying SVs, especially the smaller CNVs that were beyond the resolution of old techniques, as well as copy-neutral events such as inversions and translocations. However, there are still many problems associated with identifying SVs using NGS technology, as discussed in Study IV. As the technology and analytical methods continue to improve, some of these problems may resolve. However, it is of my personal opinion that the following cannot be neglected:

1. Collaborations among various research centres. Even as the cost for whole genome high-throughput sequencing continues to drop, routine sequencing of a large number of individuals will still remain too pricy for the majority of research centres. Collaborations will push the research at a faster pace, overcoming cost and manpower issues. Take for example the 1000 Genomes Project (www.1000genomes.org), which aims to sequence 2500 individuals, and have thus far completed the sequencing of more than 1000 individuals. Such an effort was the result of collaborations of more than 70 research groups and would definitely not have been possible by a single research centre.

2. Well-studied and standardized analysis pipelines and quality-control (QC) metrics. One of the major difficulties in comparing SVs among different studies is that all studies use different algorithms and QC metrics. With NGS technology, there are already numerous algorithms to choose from, but yet no consensus on the appropriate analysis pipeline.
3. Educating a whole new discipline of 'big data biology'. As more and more genomics data are collected, the growing need for storage, processing and analysis of the data becomes more and more apparent. Already, there is a great demand for information technology infrastructure and bioinformatics team to analyse the massive amount of data, with speculations that the costs associated with down-handling, storing and analysis of the data could be more than the production of the data. Hence, we need to train new scientists to handle these upcoming challenges.
4. Beyond discovery studies. Many early works on population wide SVs are 'discovery' studies where SVs in a population are characterized. As our understanding of SVs continues to increase, we should look beyond 'discovery', but aim to collect phenotype data for association studies.
5. Integrated knowledge with RNAseq, transcriptome, proteomics etc. We still do not have a good understanding of the function of SVs in the context of human phenotypes. The integrated knowledge of SVs with transcriptome and proteomics will enhance our ability to interpret the genome.

ACKNOWLEDGEMENTS

Needless to say, the first thanks go out to my supervisors **Chia Kee Seng**, **Yudi Pawitan** and **Agus Salim**, without which this thesis would not have been possible. I sincerely thank them for their patient supervision and support throughout these 4 years and for seeing opportunities in every difficulty we faced. Thank you also for being very flexible supervisors, for giving me the opportunity and privilege to pursue a joint degree as well as attend numerous courses and conferences overseas.

Thank you to my co-authors, **Stefano Calza**, **Ku Chee Seng**, **Vikrant Kumar**, **Anbupalam Thalamuthu**, **Mark Seielstad** and **Nasheen Naidoo**, without which the publications would not have been possible. Special thanks to **Ku Chee Seng** for always patiently answering my numerous questions and for being a fantastic 'walking encyclopedia' on genotyping technologies.

To my mentor **Marie Reilly**, thank you for your care and concern, be it with regards to my academic work or my general well-being, and (together with Yudi), for all the lovely invites to your house and all those yummy dinner treats!

To my undergraduate thesis advisor **Yap Von Bing** who first introduced me to the world of genetics and genomics.

Having been fortunate to pursue my PhD both in Singapore and in Sweden, I would like to give my gratitude to friends and colleagues from the Saw Swee Hock School of Public Health at NUS, as well as from the Department of Medical Epidemiology and Biostatistics at KI.

Special mentions from NUS include **Yang Qian, Suo Chen, Teo Yik Ying, Tai Bee Choo, Sim Xue Ling, Lim Gek Hsiang, Sharon Wee, Kaavya Narasimhalu**, **Tan Chuen Seng, Gao He, Moira Khaw, Katherine Kasiman** and **Salome Rebello**.

From KI, I would like to especially thank **Li Jingmei, Hatef Darabi, Andrea Ganna, Emil Rehnberg, Myeongjee Lee, Tong Gong** and **Ting Zhuang.** To everyone else in MEB, thank you for making MEB such a delightful place to work in!

To my amazing parents and brother, thank you for always believing in me and supporting me in whatever I choose to do.

To all my wonderful friends, especially from RGS symphonic band, RJC ODAC and NUS climbing, you know who you are, thank you for all the fun times!

Last but not least, I would like to acknowledge financial support from the National University of Singapore Graduate School of Science and Engineering (NGSS) scholarship.

REFERENCES

1. Abyzov A *et al.* (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research* **21**: 974-984.
2. Alkan C *et al.* (2011) Genome structural variation discovery and genotyping. *Nature Review Genetics* **12**: 363-376.
3. Alkan C *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41**:1061-1067.
4. Aquilante CL *et al.* (2006). Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clinical Pharmacology & Therapeutics* **79**: 291–302.
5. Beroukhim R *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proceedings of the National Academy of Sciences USA* **104**: 20007-20012.
6. Bersaglieri T *et al.* (2004). Genetic Signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**: 1111-1120.
7. Chen K *et al.* (2009) BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Naure Methods* **6**: 677–681.
8. Chiang DY, McCarroll SA (2009) Mapping duplicated sequences. *Nature Biotechnology* **27**: 1001-1002.
9. Conrad DF *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.
10. Curtis D *et al.* (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Annals of Human Genetics* **72**: 261–278.
11. Dalca AV, Brudno M (2010) Genome variation discovery with high-throughput sequencing data. *Briefings in Bioinformatics* **11**: 3-14.

12. Diskin SJ *et al.* (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research* **16**:1149-1158.

13. Eid J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

14. Gibson J *et al.* (2006) Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* **15**, 789-795.

15. Handsaker RE *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on population scale. *Nature Genetics* **43**: 269-276.

16. Harrington DJ *et al.* (2005) Pharmacodynamic resistance to warfarin associated with a Val66Met substitution in vitamin K epoxide reductase complex subunit 1. *Thrombosis and Haemostasis* **93**: 23-26.

17. Hastings PJ *et al.* (2009) Mechinsms of change in gene copy number. *Nature Review Genetics* **10**: 551-564.

18. He D *et al.* (2011) Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics* **27**: 1513-1520.

19. Hert DG *et al.* (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29**: 4618-4626.

20. Huang J *et al.* (2007) Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics* **23:** 2463–2469.

21. Iafrate AJ *et al.* (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949-951.

22. Kidd JM *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**:56-64.

23. Knudsen B *et al.* (2010) A computer simulator for assessing different challenges and strategies of de novo sequence assembly. *Genes* **1**: 263-282.

24. Ku CS *et al.* (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics* **55**: 403-415.

25. Ku CS *et al.* (2011) Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* **129**:1-15.

26. Lencz T *et al.* (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences USA* **104**: 19942-19947.

27. Li LH *et al.* (2006) Long contiguous stretches of homozygosity in the human genome. *Human Mutation* **27**: 1115-1121.

28. Li H *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**:1851-1858.

29. Locke DP *et al.* (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics* **79**: 275-290.

30. Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genetics* **24**:133–141.

31. McCarroll SA *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166-1174.

32. McQuillan R *et al.* (2008) Runs of homozygosity in European populations. *American Journal of Human Genetics* **83**: 359-372.

33. Medvedev P *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13-20.

34. Medvedev P *et al.* (2010) Detecting copy number variation with mated short reads. *Genome Research* **20**: 1613-1622.

35. Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews* **11**:31-46.

36. Mills RE *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59-65.

37. Nalls MA *et al.* (2009) Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10**: 183-190.

38. Nothnagel M *et al.* (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Human Molecular Genetics* **19**: 2927-2935.

39. O'Dushlaine CT *et al.* (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. *European Journal of Human Genetics* **18**: 1248-1254.

40. Olshen AB *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.

41. Pang AW *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11**:R52.

42. Pinto *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology* **29**: 512-520.

43. Pop M *et al.* (2004) Comparative genome assembly. *Briefings in Bioinformatics* **5**: 237-248.

44. Sabeti PC *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**.

45. Sanger F *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**: 5463–5467.

46. Schadt EE *et al.* (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**: R227-240.

47. Schatz MC *et al.* (2010) Assembly of large genomes using second generation sequencing. *Genome Research* **20**: 1165-1173.

48. Scherer SW *et al.* (2007) Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**: S7-15.

49. Schloss JA (2008) How to get genomes at one ten-thousandth the cost. *Nature Biotechnology* **26**: 1113-1115.

50. Sebat J *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525-528.

51. Sudmant PH *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**: 641 – 646.

52. Teo SM *et al.* (2010) Identification of recurrent regions of copy number variation across multiple individuals. *BMC Bioinformatics* **11**:147.

53. Teo SM *et al.* (2011) Multi-platform Segmentation for joint detection of copy number variants. *Bioinformatics* **27**:11.

54. Teo SM *et al.* (2012) Regions of homozygosity in three Southeast Asian populations. *Journal of Human Genetics* **57**: 101-108.

55. Teo YY *et al.* (2009) Genome-wide comparisons of variation in linkage disequilibrium. *Genome Research* **19**: 1849-1860.

56. Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Review Genetics* **13**:36-46.

57. Venter JC *et al.* (2001) The sequence of the human genome. *Science* **291**: 1304 – 1351.

58. Voight BF *et al.* (2006) A map of positive selection in the human genome. *PLoS Biol*ogy **4**: e72.

59. Wain LV *et al.* (2009) Genomic copy number variation, human health, and disease. *Lancet* **374**: 340-350.

60. Wang K *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**:1665-167.

61. Wellcome Trust Case-Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**:713-720.

62. Winchester L *et al.* (2009) Comparing CNV detection methods for SNP arrays. *Briefings in Functional Genomics and Proteomics* **8**: 353-366.

63. Yang TL *et al.* (2010) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *The Journal of Clinical Endocrinology & Metabolism* **95**: 3777-3782.

64. Ye K *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865-2871.

65. Yoon S *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**:1586-1592.

66. Zhang NR *et al.* (2010) Joint estimation of DNAcopy number from multiple platforms. *Bioinformatics* **26**: 153-160.

67. Zhu Y *et al*. (2007) Estimation of Wafarin Maintenance Dose Based on VKORC1 (-1639 G>A) and CYP2C9 Genotypes. *Clinical Chemistry* **53**: 1199-1205.

Mei et al. BMC Bioinformatics 2010, 11:147
http://www.biomedcentral.com/1471-2105/11/147

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Identification of recurrent regions of copy-number variants across multiple individuals

Teo Shu Mei[1,2,5], Agus Salim[1,2], Stefano Calza[3], Ku Chee Seng[2], Chia Kee Seng[1,2], Yudi Pawitan[4*]

Abstract

Background: Algorithms and software for CNV detection have been developed, but they detect the CNV regions sample-by-sample with individual-specific breakpoints, while common CNV regions are likely to occur at the same genomic locations across different individuals in a homogenous population. Current algorithms to detect common CNV regions do not account for the varying reliability of the individual CNVs, typically reported as confidence scores by SNP-based CNV detection algorithms. General methodologies for identifying these recurrent regions, especially those directed at SNP arrays, are still needed.

Results: In this paper, we describe two new approaches for identifying common CNV regions based on (i) the frequency of occurrence of reliable CNVs, where reliability is determined by high confidence scores, and (ii) a weighted frequency of occurrence of CNVs, where the weights are determined by the confidence scores. In addition, motivated by the fact that we often observe partially overlapping CNV regions as a mixture of two or more distinct subregions, regions identified using the two approaches can be fine-tuned to smaller sub-regions using a clustering algorithm. We compared the performance of the methods with sequencing-based results in terms of discordance rates, rates of departure from Hardy-Weinberg equilibrium (HWE) and average frequency and size of the identified regions. The discordance rates as well as the rates of departure from HWE decrease when we select CNVs with higher confidence scores. We also performed comparisons with two previously published methods, STAC and GISTIC, and showed that the methods we consider are better at identifying low-frequency but high-confidence CNV regions.

Conclusions: The proposed methods for identifying common CNV regions in multiple individuals perform well compared to existing methods. The identified common regions can be used for downstream analyses such as group comparisons in association studies.

Background

Copy-number variants (CNVs) are genomic regions that contain an abnormal number of copies. In humans, we normally expect two copies of each autosomal region, but in CNV regions we may observe copy gains or losses. Current common technology used for CNV detection are high-density single nucleotide polymorphism (SNP) arrays or array comparative genomic hybridization (aCGH) arrays. Detection of CNVs from aCGH arrays is mostly based on locating change-points in intensity-ratio patterns that would partition each chromosome into several discrete segments [1-5]. On the other hand, the hidden Markov model (HMM) is particularly popular for detection of CNVs from SNP arrays, where the hidden states provide a natural way of combining information from the total signal intensity and the allele frequency values (see for example, [6,7]). These approaches detect CNVs sample-by-sample, and because of the high noise level in the intensity values, especially for SNP array data, the boundaries of the detected CNVs tend to vary among individuals. However, in a homogenous population, common CNV regions are likely to occur at the same genomic locations across different individuals. Our focus in this paper is to identify common CNV regions in multiple individuals from a given population.

Common CNV detection algorithms for SNP arrays report the log Bayes factor as a confidence score for each identified region; this provides a measure of the

* Correspondence: yudi.pawitan@ki.se
[4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, Stockholm 17177, Sweden

reliability of a detected CNV within an individual. Previous methods developed to identify recurrent CNV regions (see [8] for a review) were primarily developed for aCGH data and hence did not incorporate confidence scores. For example, a previously published method, STAC [9], uses two statistics to identify recurrent CNV regions. These statistics are based on the frequency of occurrence of the regions and the alignment of the regions. However, since the method does not incorporate confidence scores, every individual region contributes equally to the statistic, whereas in fact, inter-sample variability is bound to exist, where some regions are more likely to be true/false positives. Furthermore, STAC requires each chromosome to be split into non-overlapping windows of a user-defined fixed size. The algorithm then searches for evidence of common CNV regions within each window. The weakness of this is that the output from such an approach will only provide evidence of whether each window harbours a common CNV, but will not indicate the breakpoints of the CNV. Although we may decrease the window size to improve the resolution, in practice, doing so will incur an enormous computational burden.

In this paper, we investigated two different methods to detect common CNV regions. The methods take segmented data as the input. The first method estimates a statistic based on the frequency of occurrence of reliable CNVs, where reliability is determined by a high confidence score. The second method is based on a weighted frequency of occurrence of CNVs, where the weights are determined by the confidence scores. Figure 1 illustrates a common CNV region in chromosome 22, identified using the first method, and shows evidence of several distinct subregions within the identified common region. Hence, in addition to these methods, we also investigated the use of a clustering algorithm to split the common regions into smaller subregions.

To assess the performance of the methods, we ran the algorithms on 112 HapMap samples from the Illumina iControl database, composed of individuals from three populations (Yoruba, Caucasian and Asian). We compared the regions we identified to the regions identified using sequencing [10]. In general, the discordance rates with sequencing-based CNV regions as well as the rates of departure from HWE decreased when we filtered the individuals with a stricter confidence score threshold. To benchmark the proposed methods to currently available methods, we performed comparisons with STAC [9] and GISTIC [11] and found that the proposed methods outperformed both STAC and GISTIC in identifying low-frequency but high-confidence CNV regions.

Methods

Data Structure

We assume that the raw intensity data have been processed by a CNV detection algorithm. Denote by $R_i = \{R_{i1}, R_{i2}..., R_i = \{R_{i1}, R_{i2}, \ldots, R_{i\ell_i}\}\}$ the collection of CNV regions detected in individual i, for $i = 1,...,n$. A region is defined by its start and end probe locations, and its CNV type (duplication or deletion). For each region, we assume we have a confidence score statistic that measures the likelihood that the detected region is real. An example of this statistic is the log Bayes Factor (see [6]). For region j detected in individual i, we denote this statistic as C_{ij}.

Cumulative Overlap Using Very Reliable Regions (COVER)

Our confidence in a CNV region depends on the within- and between-subject information; our methods shall utilize both information. The within-subject information comes from the strength of the signal within an individual CNV region, and this is measured by the confidence score. The between-subject information comes from the consistency of the CNVs across different individuals. Intuitively, we have less confidence in a CNV that occurs in one individual than one that occurs in many individuals. However, a single occurrence of CNV might still be a true discovery if it is associated with a high confidence score, i.e., it is based on a strong signal.

Since individual CNV regions span different probes, the number of individual regions that overlap each probe varies. However, common CNV regions tend to occur at almost the same genomic locations across multiple individuals. Hence, we expect the common regions to be identified by consecutive probes where a 'significant' number of individuals have an overlapping CNV region. Furthermore, we also expect the confidence score of the individual region to be relatively high.

Let Z_{ijk} be the indicator that region j detected in individual i overlaps with probe k. For each probe k, we calculate the Cumulative Overlap using Very Reliable Regions (COVER) statistic y_k, defined as

$$y_k = \sum_{i=1}^{n} \sum_{j=1}^{\ell_i} (Z_{ijk} \times I_{C_{ij} \geq c}),$$

where $I_{C_{ij}>c}$ is the indicator function for regions detected with a confidence score above a certain threshold c. The common CNV regions are then defined by

$$\Re = \left\{ \left[l_m, l_{m'} \right], y_k \geq u, \forall_k \in \left[m, m' \right] \right\},$$

representing sets of consecutive probes for which y_k is consistently greater than or equal to a specified

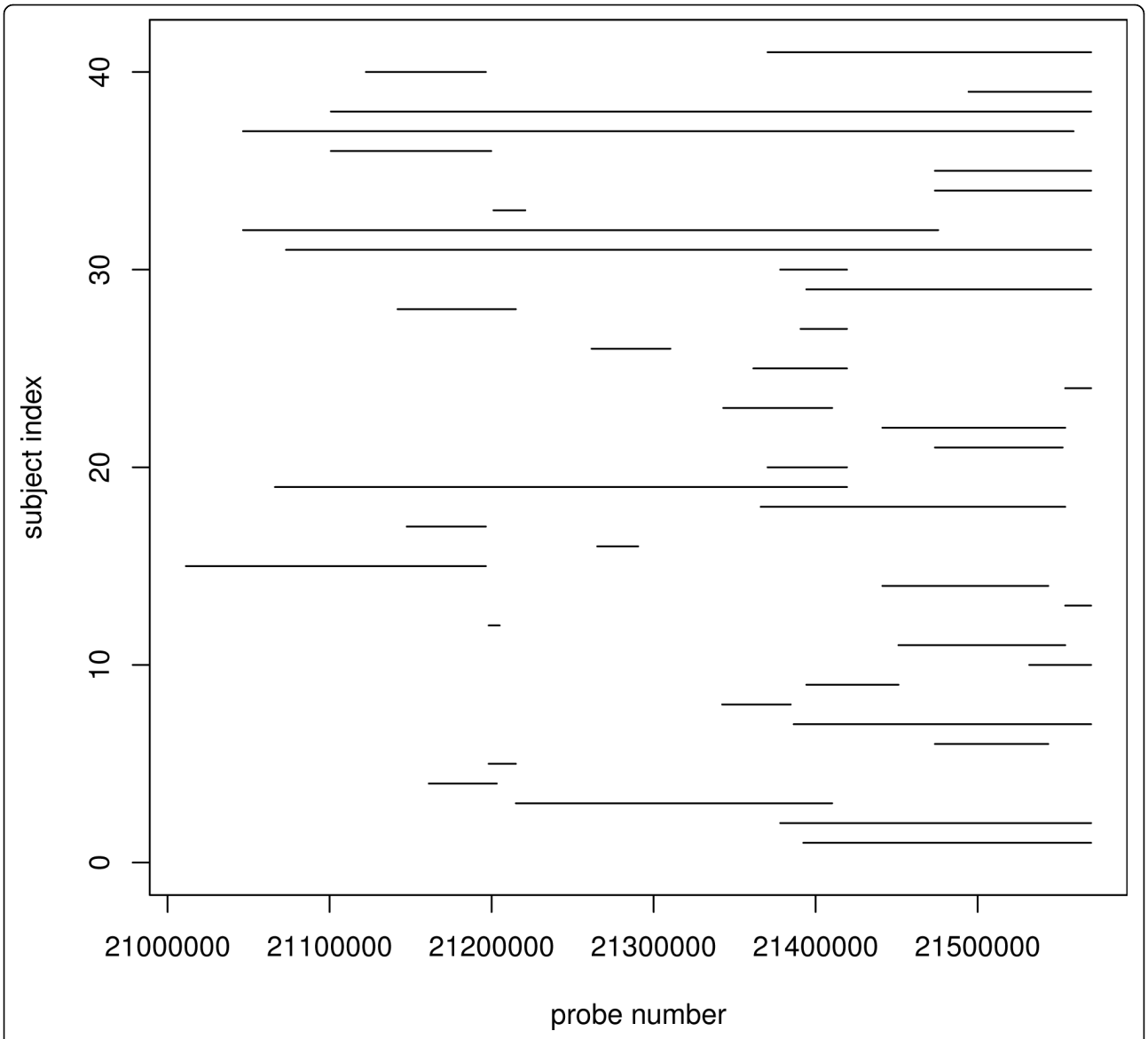


Figure 1 An example of a common CNV region found based on COVER method with threshold $u = 2$ and $c = 60$. This figure illustrates a common CNV region in part of chromosome 22, found using the COVER method with threshold $u = 2$ and confidence cutoff at 60th percentile. 41 out of 112 individuals have CNVs that overlap with this common region, indicated by the horizontal lines. We can see that despite being identified as a common region, the individual regions still portray a mixture phenomenon of several distinct subregions.

threshold u. l_m is the genomic position of probe m and it is implicitly understood that the cardinal position of the probe reflects its relative position in the chromosome so that when there are M probes in a chromosome, $l_1 < l_2 < ... < l_M$.

Using COVER, we can identify multiple common CNV regions within a chromosome. Furthermore, different subsets of individuals may contribute to different common regions, hence allowing COVER to identify regions that are common to only a subset of individuals. By only considering individual regions that are detected with high reliability, we also incorporate the uncertainty associated with each individual region in the identification of common regions. If this is not taken into account, then all regions would be treated equally despite the fact that some are more likely to be true than the others. Figure S4 in the [Additional File 1] gives an illustration of how COVER works.

Cumulative Composite Confidence Scores (COMPOSITE)

In COVER, regions with low confidence are given zero weights and they do not contribute to the COVER statistic. The within-subject confidence is not fully exploited when computing the COVER statistic: regions

that are detected with low confidence but nonetheless detected consistently across a large number of subjects might be missed.

This limitation is addressed in the second method. For probe k the composite confidence score (COMPOSITE) statistic is defined as,

$$s_k = \sum_{i=1}^{n} \sum_{j=1}^{\ell_i} (Z_{ijk} \times C_{ij}).$$

This formula is in fact similar to COVER statistic, where instead of using the indicator function $I_{C_{ij}>c}$ as weights, now all detected individual regions contribute to the COMPOSITE statistic, with the amount of their contribution proportional to their confidence scores.

Using COMPOSITE, the common CNV regions are then defined as

$$\mathfrak{R} = \left\{ \left[l_m, l'_m \right], s_k \geq v, \forall_k \in \left[m, m' \right] \right\},$$

representing sets of consecutive probes for which s_k is consistently greater than or equal to a specified threshold v. Figure S4 in [Additional file 1] gives an illustration of how COMPOSITE works.

Clustering of Individual CNV Regions within a Common Region (CLUSTER)

Cluster analysis has been used in the analysis of gene expression and aCGH data (see for example, [12-14]). Here, the motivation for CLUSTER stems from the observation that within a common CNV region identified by COVER or COMPOSITE, a complex mixture phenomenon can still be observed (see Figure 1).

Figure 2 depicts the hypothetical situation where a common region of length L bases has been identified by COVER or COMPOSITE. Four individual regions overlap with the common region and from the figure, it is clear that the first two regions are clustered to the left while the last two are clustered to the right. The two groups may form two distinct subregions and these

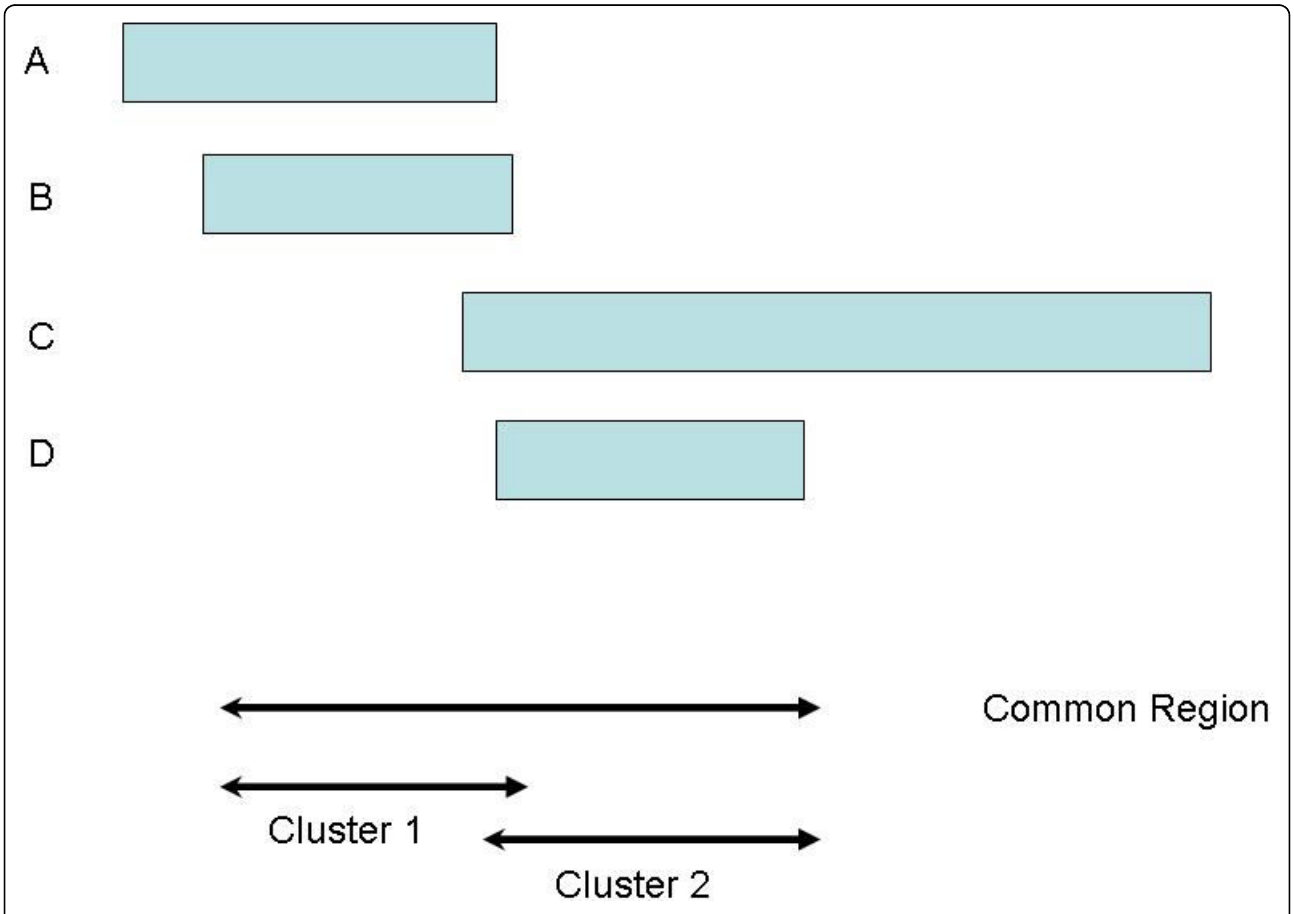


Figure 2 Hypothetical example of a identified common CNV region with 2 distinct clusters. Hypothetical situation where an identified common CNV region is common to four individuals. From the figure, it is clear that the common region consists of two partially overlapping regions. The first two individual regions are clustered together to the left of the common region, while the last two individual regions are clustered to the right.

subregions could differ biologically. In reality, the situation is more complex than the hypothetical example here (see for example Figure 1).

To find the subregions inside this common region, we first perform pairwise comparisons of the individual regions that overlap with the common region. For example, the comparison of two regions A and B can be summarized into 4 values (a, b, c, d), where a is the number of bases for which both A and B overlap with the common region, b is the number of bases where A overlaps with the common region but B does not, c is the number of bases where B overlaps with the common region but A does not, and $d = L - a - b - c$.

The (dis)similarity index can be computed using a variety of distance metrics appropriate for binary data such as the Manhattan, Canberra or Jaccard distance [15]. The Jaccard distance is particularly attractive for our case; it is defined by $a/(a + b + c)$ and can be interpreted as the percentage of common overlap of the two regions relative to the union of the overlaps of the two regions with the common region. We then construct a dissimilarity matrix as input to a hierarchical clustering algorithm. The number of clusters will be determined by the amount of within-cluster similarity we require. The boundaries of each subregion will be the minimum and maximum positions of all individual regions that belong in that cluster. If these bounds overshoot the boundaries of the initially identified region, then the boundaries will be reset to the boundaries of the initial region.

Results and Discussion

Assessment and Comparison

Datasets

We studied the performance of the proposed procedures by varying the corresponding threshold parameters in each approach. 112 HapMap samples, comprising 46 Caucasian (CEU), 29 Beijing Chinese and Tokyo Japanese (CHBJPT) and 37 Yoruban (YRI) individuals were used in the analysis. These samples are part of the Illumina iControl Database. Each sample was genotyped using the Illumina 1M chip, and PennCNV [6] was used to detect the individual CNV regions.

Comparison with Sequenced Regions

We compared the common regions we identified to a list of reference CNVs identified in eight HapMap samples using sequencing data [10]. For each of the eight samples, we calculated the discordance rates by recording the proportion of common CNV regions (found using our methods) for that sample that were not concordant with the sample-specific reference CNVs. To be 'concordant' with a reference CNV, a region has to be either contained within the reference CNV or it has to overlap with at least 50% of the reference region. It is important to note however that it is difficult to get a gold standard for common CNV boundaries; even the sequencing-based CNV regions cannot be expected to have 100% sensitivity and specificity in genotype calling and certainly not in boundary calls for common CNVs.

Comparison with other Array-based Regions

We compared the regions found using our methods to the regions found by two other groups using array-based methods. We compared with McCarroll *et al*. [16], where the regions were identified using the Affymetrix SNP 6.0 arrays on 270 HapMap samples. To minimize false discoveries, they ran two independent experiments and require a CNV to be observed in both experiments. We also compared our regions to the regions found by Conrad *et al*. [17]. These regions were identified using tiling oligonucleotide microarrays, comprising of 42 million probes, on 41 HapMap samples. A total of 11,700 CNVs were identified, and 8,599 were validated using a set of stringent criteria including (i) additional measurements by Agilent 105K CGH arrays, (ii) overlap with previous studies and (iii) other quality-control filters. For our comparisons, we used only the 8,343 validated CNVs in the autosomal regions.

Comparison to other approaches

We compared our approaches to previous common CNV detection methods, STAC: Significance Testing for Aberrant Copy number [9] and GISTIC: Genomic Identification of Significant Targets in Cancer [11].

Briefly, STAC takes segmented data as input and estimates two statistics: 1. A frequency statistic, which estimates the frequency of aberration at each location across all individuals. 2. A footprint statistic, which uses a subset search methodology and counts the number of locations c such that c is contained in a set of intervals (see [9] for more details). It then uses a permutation test to assess the significance of the observed region. STAC requires each chromosome to be split into non-overlapping regions of a user-defined fixed size. The algorithm looks for evidence of common CNV regions within each window, and reports the associated frequency and footprint p-values.

GISTIC first calculates a 'G score' that is associated with both the frequency of occurrence as well as the amplitude of the aberration. Then, it calculates the probability (q-value) of the observed region occurring by chance via a permutation test. One can either input the log intensity ratios, where the GLAD algorithm [18] will be used to segment the data, or input pre-segmented data using other algorithms.

We had also planned to make comparison to another method called MSA [19], but failed because the software, which is part of the GenePattern module, did not work properly. MSA can be viewed as an improvement over STAC, where it extends the notions of frequency

and footprint statistics using original intensity ratio data instead of segmented data [8]. We also tried a comparison to RJaCGH [2], which uses a non-homogenous Hidden Markov Model fitted via the Reversible-Jump Markov Chain Monte Carlo method to estimate the probability that a region has copy number alterations; the method also allows the identification of minimal common regions of copy number changes among multiple individuals.

Unfortunately, with our samples, the algorithm did not converge, so we could not proceed with the comparison.

Testing Hardy-Weinberg Equilibrium

It has been observed that the majority of common CNV regions are inherited [20]. Hence, for a population of normal (healthy) individuals, we expect, for most of the common regions, the integer copy numbers to be in Hardy-Weinberg equilibrium (HWE). The small number of regions that depart from HWE can be attributed to factors such as recent mutations. For example, McCarroll *et al*. [16] found that about 98% of common diallelic CNV regions do not show significant departure from HWE. In principle, HWE applies to both diallelic CNVs (where only loss or gain of copy numbers are present in addition to normal copies) and multi-allelic CNV regions (where both loss and gain of copies are present).

For diallelic CNVs with only loss and normal-copy numbers (copy-number = 0,1,2), the HWE test can be conducted by treating '0' copies as minor allele homozygous, '1' copy as heterozygous and '2' copies as reference homozygous. Similarly, for CNVs with only gain and normal-copy numbers (copy-number = 2,3,4), we treat '2' copies as reference homozygous, '3' copies as heterozygous and '4' copies as minor-allele homozygous. For multi-allelic CNVs, a model with three or more alleles is needed. However, the HWE test cannot be performed directly on the unphased copy-number because there is an issue with different combinations of alleles producing the same copy-number. For example, in a 3-allele model, a copy-number of 2 can be produced by a combination of '0' and '2' copies or two '1' copy alleles.

When dealing with samples from healthy individuals, we propose to use the outcome of the HWE tests to select 'optimal' parameter thresholds (e.g., c in COVER and v in COMPOSITE). If we observe a large number of common CNV regions with significant departure from HWE (after accounting for population stratification), it could mean that the parameters we choose are not optimal. When dealing with a mixture of healthy and diseased individuals such as in association studies, it is expected that the CNVs among the diseased individuals will show some degree of departure from HWE as some of the CNVs could be due to recent abberations. We propose performing HWE tests only among the healthy individuals to select the optimal threshold parameters.

Results

COVER results

Figure 3 shows the results for COVER. The discordance rates with Kidd *et al*.'s [10] reference CNVs (see Comparison with Sequencing Results) can be as high as 80% when we include all CNV calls in identifying the common regions. The discordance rates decrease when we exclude CNVs whose confidence scores are below a certain percentile; more severe filtering generally reduces the discordance rates. The lowest discordance rates of about 55% were achieved when we excluded individual regions whose confidence scores were below the 80th percentile. Surprisingly, increasing the required minimum number of individuals inside a region (u) does not seem to have an effect on the discordance rates.

However, the required minimum number of individuals (u) does affect the rates of HWE violation (calculated as the percentage of diallelic CNVs whose p-value from the HWE test is < 0.01 in at least one of the three ethnic groups). (Some HapMap individuals were related; the HWE test in each ethnic group was carried out on unrelated individuals only.) There is an overall increasing trend for the proportion of common CNV regions that violate HWE when we increase the minimum number of individuals (Figure 3(b)). This is partly due to the fact that with increasing number of individuals, we detect CNV regions with larger minor allele frequencies (see Figure 3(c)), hence the test for HWE will be more powerful. Generally, the rates of departure from HWE are less than 10% and can be lowered by filtering out individuals with lower quality regions. A steeper reduction in the rates of departure from HWE can be observed when only individual regions whose confidence scores are above the 60th percentile are considered (Figure 3(b)).

The sizes of the identified common regions generally increase when we filter lower quality individual regions (Figure 3(d)), reflecting the fact that smaller regions with fewer overlapping probes would tend to have lower confidence scores. By choosing confidence score thresholds (c) anywhere up to the 60th percentile, the average size of the common regions are approximately the same or slightly smaller than the average size that Kidd *et al*. [10] obtained using sequencing methods (solid horizontal line in Figure 3(d)). The dashed horizontal line in Figure 3(d) shows that the median size of CNV regions identified using the 500K EA chip [21] is much larger than what we observe using our methods.

For this dataset, setting the confidence score threshold to the 60th percentile seems to be the optimum choice. With this setting, the discordance rates are around 60%

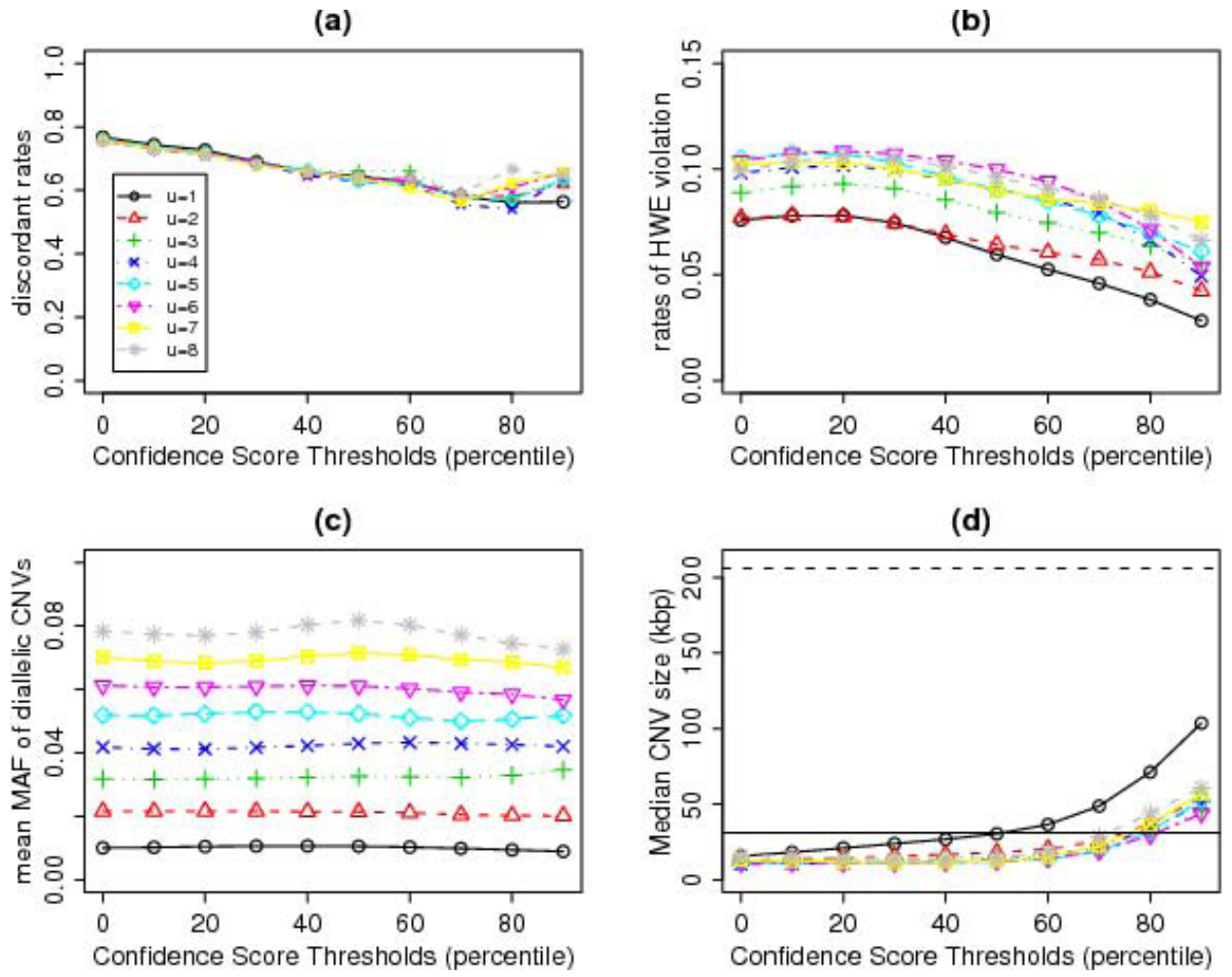


Figure 3 Results of COVER method. (a) Discordant Rates, (b) Proportion of diallelic CNVs that failed HWE, (c) mean minor allele frequency (MAF) of diallelic CNVs and (d) Mean CNVs size (kilo-bases) as a function of confidence scores cut-off points and minimum number of individuals.

and the proportion of diallelic CNVs that violate HWE is kept at around 8%. The choice of u is more subjective, as it depends on our definition of 'common' regions. For example, if we require each common region to overlap with at least three individual regions and set c to the 60th percentile, we will identify 443 common CNV regions (see [Additional file 2]).

COMPOSITE results

A total of 89% of the probes does not contain any individual CNV regions and thus their composite scores are zero. So, if we set the threshold v at the 89th percentile of the composite scores, we do not filter out any individual regions and this approach is essentially the same as using $u = 1$ and $c = 0$ in COVER.

Figures 4(a) and 4(b) show that, as we increase the threshold, the discordance rates as well as the rates of HWE violation decrease steadily. Unlike the COVER approach, where increasing the confidence score threshold does not result in lower ability to detect rarer CNVs, increasing the composite score threshold does result in fewer rare CNVs being detected (Figure 4(c)). This is because the composite score is a function of both the confidence score and the number of individuals within a common region. By increasing the threshold, we are implicitly requiring more individuals within a common region.

The increasing trend of mean minor allele frequency (MAF) is consistently seen when the threshold is increased to the 96th percentile. Beyond this, the mean MAF decreases because large regions with higher MAF may be split into several subregions with smaller MAF. This observation is consistent with the pattern of median size of CNV regions (Figure 4(d)). Generally, we are losing the smaller regions with low composite scores as we increase the threshold. However, beyond the 96th

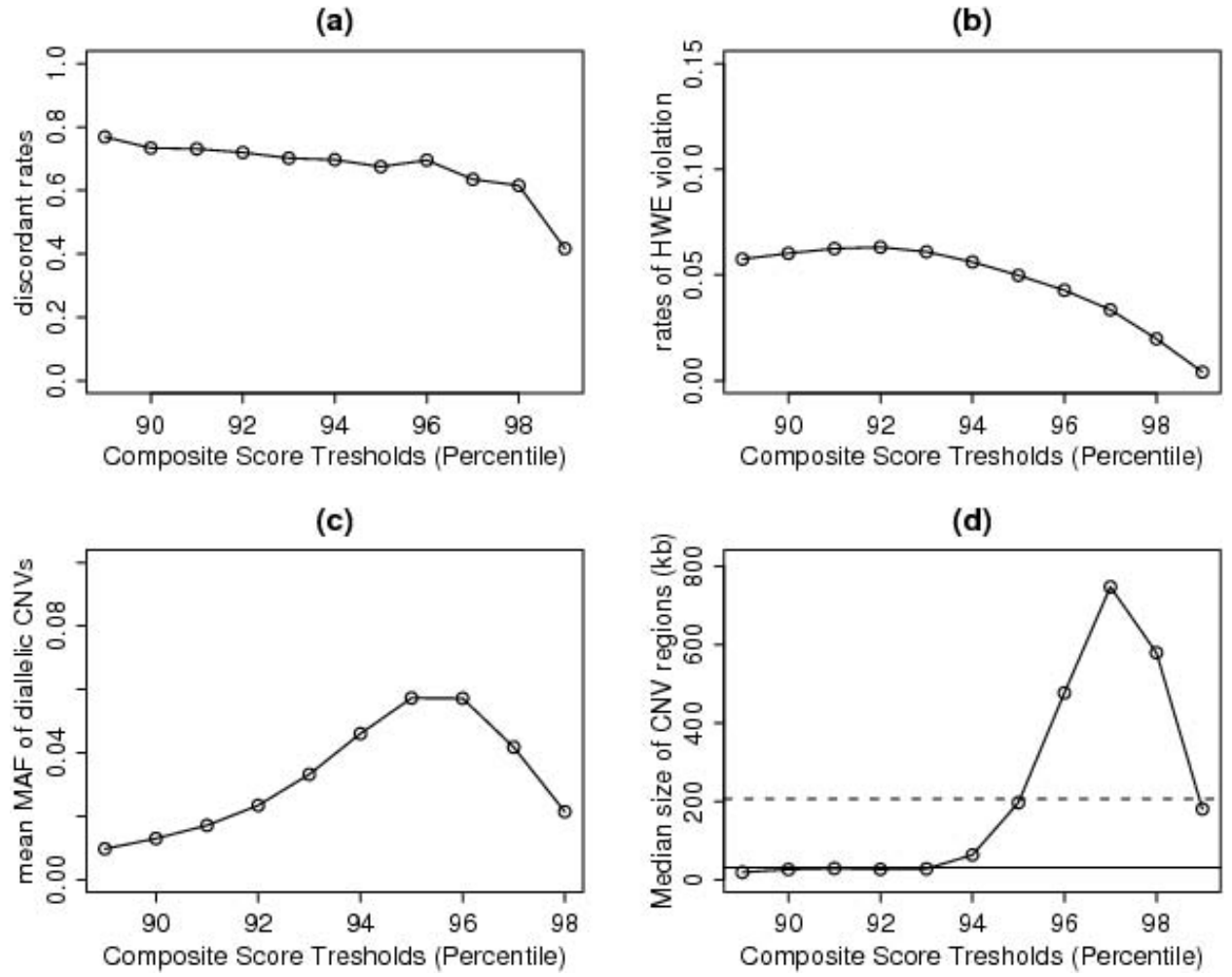


Figure 4 Results of COMPOSITE method. (a) Discordant Rates, (b) Proportion of diallelic CNVs that failed HWE, (c) mean minor allele frequency (MAF) of diallelic CNVs and (d) Median size of CNV regions (kb) as a function of composite confidence scores cut-off points. Solid line is median CNV size found by Kidd *et al*.

percentile, the median region size decreases again due to the splitting of the large regions.

The optimal setting is to set the threshold to the 94th percentile, where the proportion of regions that failed HWE is around 5% (Figure 4(c)). Using this setting, we are able to detect 491 CNV regions (see [Additional file 3]) with median CNV size slightly larger then the median size found by Kidd *et al*. [10]. The discordance rates among the eight HapMap samples are approximately 70%, higher than what can be achieved by COVER. Hence, although COMPOSITE can pick up more regions, a higher percentage of these regions is likely to be false discoveries.

CLUSTER results

The common regions identified using either COVER or COMPOSITE can be further refined into distinct subregions using CLUSTER. Here, we present the results of applying CLUSTER to the common regions identified by COVER. We choose the CLUSTER parameters so that regions will be clustered together if they are at least 60% similar. Complete linkage is used so that the distance between any pair of clusters is defined as the maximum distance between a pair of members drawn one from each cluster. Single or average linkage can also be used. Since single linkage defines the distance between any pair of clusters as the minimum between a pair of members from the clusters, it generally tends to produce clusters that are more similar to each other, and when the same similarity cut-off point is used, it tends to produce fewer clusters than complete linkage. Meanwhile, using average linkage gives more clusters than single linkage, but fewer than complete linkage. In the [Additional file 1], we compare the three linkage measures for a sample region.

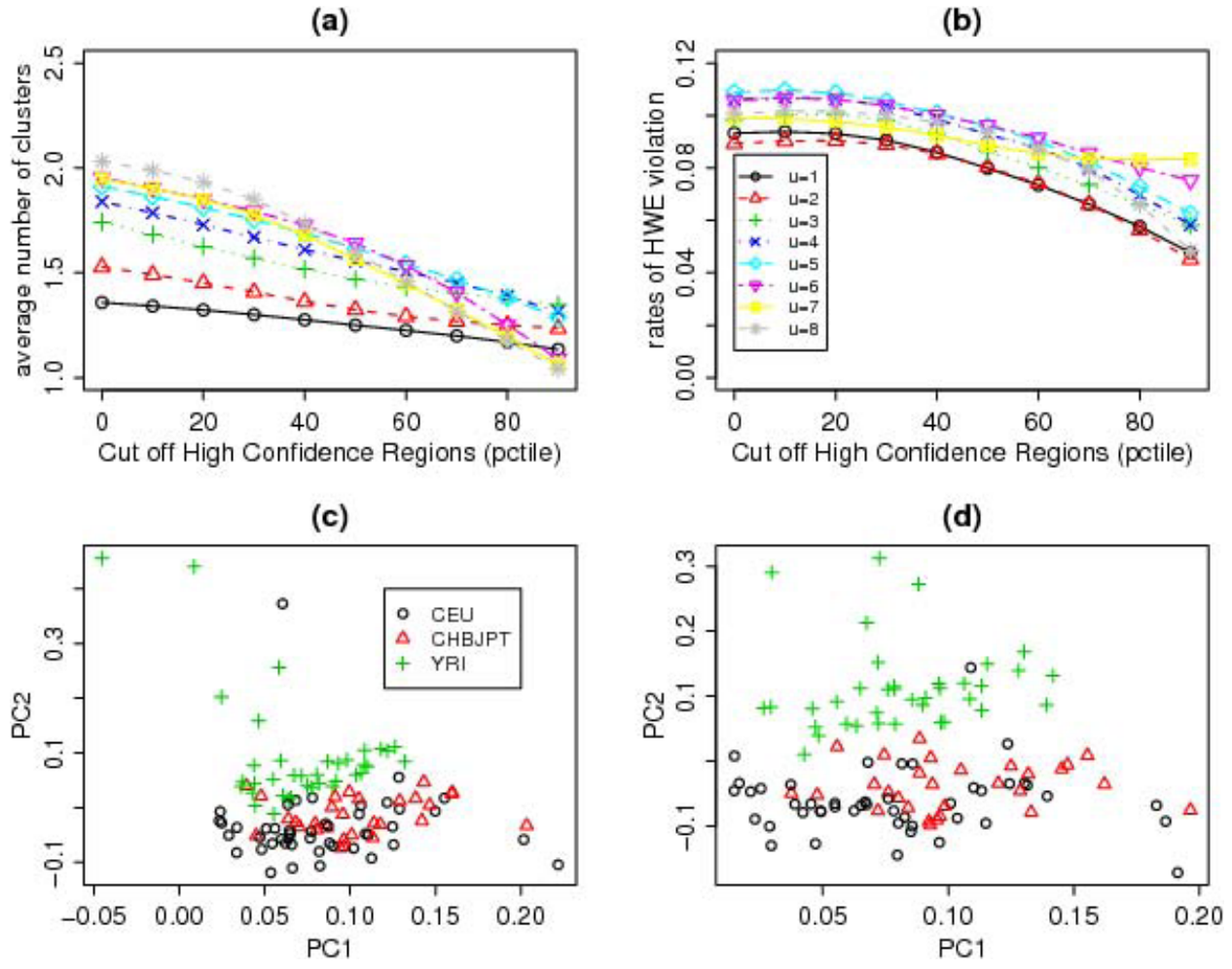


Figure 5 Results of applying CLUSTER to common regions identified by COVER method. (a) Average number of clusters, (b) rates of departure from HWE, (c) First and second components of PCA based on subjects' integer copy-number calls at common regions found using COVER (with $u = 3$ and $c = 60$), (d) First and second components of PCA based on subjects' integer copy-number calls at common regions found using complete-linkage CLUSTER (`with cluster.limit = 0.6`).

Figure 5(a) shows that the number of clusters decreases when we increase the confidence score threshold. But even when we consider CNVs with confidence scores above the median, the clustering effect is still evident with 1.3 to 1.7 clusters found for each common region, depending on which threshold value u is used. For the optimum parameters $u = 3$ and $c = 60$, on average, 1.5 clusters are found per common region. The rates of departure from HWE (Figure 5(b)) are approximately the same as in Figure 3(b) and increasing the confidence-score threshold lowers the rates.

Once the common regions are identified, it is straightforward to perform a number of downstream analyses. For example, a principal component analysis (PCA) can done based on subjects' integer copy-number calls at these regions (see Section 'Principal Component Analysis of CNV Profiles' for more details). In the HapMap dataset, CLUSTER clearly improves the separation between the Yoruba and the other two populations based on the subjects' common CNV region profiles (compare Figure 5(c) vs 5(d)). This result suggests that different ethnic groups have more subtle differences in the breakpoints of CNV regions.

Comparisons

McCarroll et al.'s versus Kidd et al.'s Results Using the Affymetrix 6.0 arrays, McCarroll *et al.* [16] employed a set of strict criteria based on duplicate experiments to identify the CNV regions. For each of the eight samples sequenced by Kidd *et al.* [10], we calculated the discordance rates with McCarroll *et al.*'s CNVs and they range from 71% for sample NA12878 to 84% for sample NA18517. On average, across the eight samples, 76% of the regions found by McCarroll *et al.* are discordant with the regions found by Kidd *et al.*

[10]. In comparison, using COVER, the discordance rates are around 60% (see Section "COVER Results"). Thus, the methods described in this paper, using only data from a non-duplicated experiment, actually perform better in terms of discordance rates against sequencing data.

McCarroll et al.'s versus COVER/COMPOSITE Results We also compared the regions identified by our approaches to the list of common CNV regions identified by McCarroll *et al*. [16]. Figure 6(a) shows that by using COVER, the discordance rates can be lowered by either increasing the confidence-score threshold, placing a higher limit on the minimum number of individuals (u), or both. For the best scenario, the discordance rate is about 15%. Using COMPOSITE, the discordance rates can be reduced by increasing the composite-score threshold, but even for the best scenario, the discordance rate is around 25% (see Figure 6(b)).

Comparison to Conrad et al.'s regions Treating the set of 8,343 validated autosomal CNVs found by Conrad et al. [17] as reference CNVs, we calculate the discordance rates against this reference list. Using the optimal parameters for COVER/COMPOSITE for this dataset, we obtain discordance rates of 42% and 31% for COVER and COMPOSITE respectively. By refining the regions using CLUSTER, the discordance rate for COVER decreases to 34% and that for COMPOSITE remains about the same, at 33%. These are better than McCarroll *et al*.'s [16] regions, which have a discordance rate of 44%.

Comparison to GISTIC As input to GISTIC, we used CNV calls from PennCNV for the same Hapmap samples as described in the Datasets Section. Using the default parameters of GISTIC, with the q-value threshold set at 0.25, we obtained 342 significant common regions with a mean frequency of 0.106 and a median confidence score of 15.7. For comparison with COVER and COMPOSITE, we chose threshold parameters to give the closest number of common regions to that detected by GISTIC. For COVER, this corresponded to the choice of u = 3 and c = 70th percentile, which yielded 329 regions with a mean frequency of 0.065 and median confidence of 32.3. For COMPOSITE, the threshold was chosen to be the 94.5th percentile, and this yielded 360 regions with a mean frequency of 0.121 and median confidence of 27.6.

For each region identified by COVER, we checked if it was concordant with any region identified by GISTIC. Concordance is defined in the same way as in the Section 'Comparison with Sequencing Results'. The COVER-identified regions can hence be divided into two groups: those that are concordant with at least one GISTIC region and those that are not. For each group, we computed the mean frequency and median confidence score, as well as the discordance rates with Kidd et al.'s regions. We did the same for each region identified by GISTIC, checking if the region was concordant with any region identified by COVER. Similar analysis was done comparing COMPOSITE and GISTIC.

Table 1, for COVER, shows that regions that are concordant with GISTIC regions have higher frequencies but moderate confidence scores, while those that are not concordant with GISTIC regions have lower frequencies but higher confidence scores. The concordant

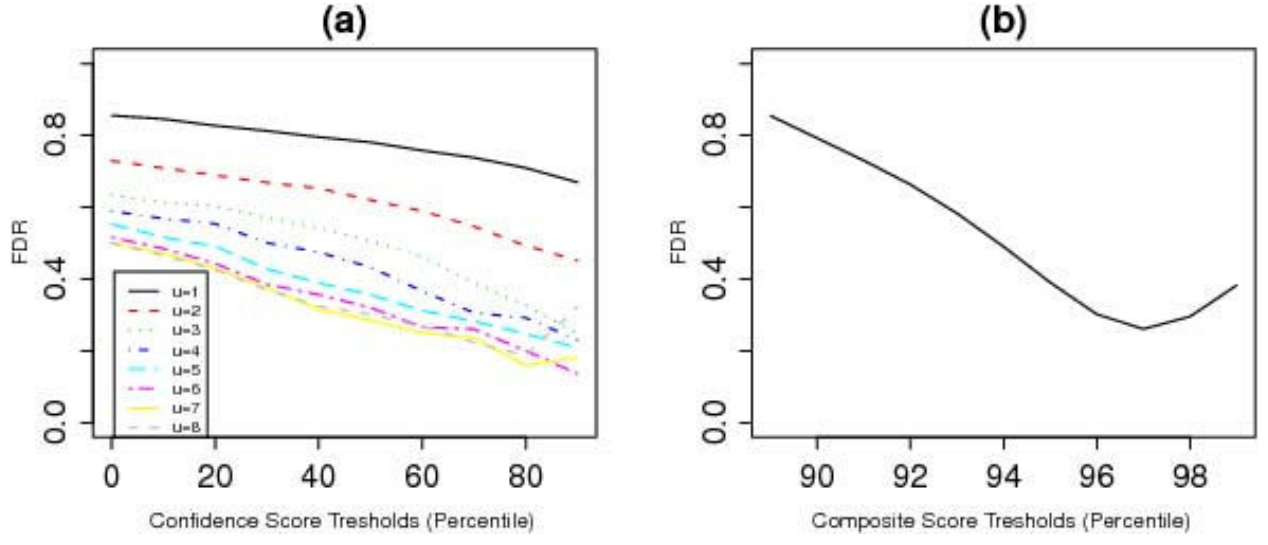


Figure 6 Comparison to McCarroll's CNVs. (a) Discordance rates when comparing regions found using COVER and those found by McCarroll *et al*., plotted against confidence score thresholds for different values of u. (b) Discordance rates when comparing regions found using COMPOSITE and those found by McCarroll *et al*., plotted against composite score thresholds.

Table 1 Comparison with GISTIC.

regions found by	overlap?	no. of regions	mean freq	median conf	discordance**
COVER	✓ GISTIC	139	0.10	30	62%
	✗ GISTIC	190	0.037	37.5	87%
COMPOSITE	✓ GISTIC	162	0.21	20.0	64%
	✗ GISTIC	198	0.048	72.8	75%
GISTIC	✓ COVER	153	0.15	22.3	56%
	✗ COVER	189	0.072	8.8	84%
	✓ COMPOSITE	173	0.15	20.6	61%
	✗ COMPOSITE	169	0.058	8.8	82%

✓ - overlap

✗ - no overlap

** discordance rates with Kidd's sequencing results.

This table shows a summary of the results obtained from comparing COVER/COMPOSITE to GISTIC.

regions have lower discordance rates with sequenced-based results. Similar patterns in frequencies, confidence scores and discordance rates are also seen for the regions found by COMPOSITE. We deduce that GISTIC misses regions that are of low frequencies but high confidence scores. Hence, it seems that COVER/COMPOSITE can identify the low-frequency CNVs better. In addition, of the regions found by GISTIC, those that are concordant with COVER or COMPOSITE have high frequencies and moderate confidence scores while those that are not concordant have low frequencies and low confidence scores. Again, the concordant regions have lower discordance rates with sequenced-based results. From this, we deduce that the regions identified by GISTIC but missed by our methods are those with low frequencies and low confidence scores, and hence more likely to be false positives.

Comparison to STAC For the purpose of analysis using STAC, we split each chromosome into 1450-1500 fixed-size windows with the size of the windows varying from 165 kb for chromosome 1 down to 24 kb for chromosome 22, resulting in a total of 32780 windows across chromosome 1-22. (We tried a smaller window size but the computational burden became too large, where even after 48 hours the algorithm was still running in a 3 GHz windows PC with 4 Gb RAM). We used 0.05 as a cut-off to declare windows with significant frequency or footprint p-values, and obtained 868 significant windows with a mean frequency of 0.155. Each significant fixed-size window will be taken as a significant region.

To compare the regions found by STAC to the regions found using COVER and COMPOSITE, we chose threshold parameters to give a number of common regions closest to that detected by STAC. For COVER, this corresponded to the choice of u = 2 and c = 60th percentile, and for COMPOSITE, the 93th percentile. We obtained 777 and 805 common regions respectively. We performed similar analysis as in the comparison to GISTIC.

A summary of this comparison is shown in Table 2a. We observe similar results as in the comparison to GISTIC: regions that were identified by STAC but that were missed by COVER/COMPOSITE have low frequencies and low confidence scores, but regions identified by COVER/COMPOSITE that were missed by STAC have low frequencies but high confidence scores, and were thus more likely to be true positives.

We also investigated if the relative performance of STAC would improve if we manually filtered out individual regions with lower confidence scores. We decided to use only individual regions whose confidence scores were above the median confidence score of all reported regions. Using this filtered input, STAC identified 654 significant windows. Using u = 2 and c = 70th percentile for COVER and the 93.5th percentile for COMPOSITE, we identified a similar number of common regions (615 for COVER and 610 for COMPOSITE). Table 2b summarizes the results of this comparison and our conclusions are similar to those with the unfiltered input data.

We conclude that COVER and COMPOSITE are able to detect the majority of the regions found by STAC, and in addition they also detect common high-confidence CNV regions that occur in a smaller number of subjects that were missed by STAC.

Implementation

The methods are implemented in an R package `cnvpack`. The main input is a list of detected individual CNV regions with the following information: Sample name, chromosome number, detected integer copy number, start and end genomic locations and a confidence score. The package can be downloaded from http://www.meb.ki.se/~yudpaw.

Table 2 Comparison with STAC.

STAC input: all data regions found by	overlap?	no. of regions	mean(freq)	median(conf)
COVER	✓ STAC	301	0.084	25.6
	✗ STAC	476	0.021	31.2
COMPOSITE	✓ STAC	372	0.14	18.6
	✗ STAC	433	0.023	52.5
STAC	✓ COVER	609	0.15	23
	✗ COVER	259	0.11	8.1
	✓ COMPOSITE	727	0.15	20.5
	✗ COMPOSITE	141	0.07	7.21
STAC input: filtered data regions found by	overlap?	no. of regions	mean(freq)	median(conf)
COVER	✓ STAC	294	0.068	30.2
	✗ STAC	321	0.020	37.6
COMPOSITE	✓ STAC	297	0.14	23.1
	✗ STAC	313	0.045	65.2
STAC	✓ COVER	585	0.14	28.1
	✗ COVER	69	0.07	16.1
	✓ COMPOSITE	595	0.14	26.8
	✗ COMPOSITE	59	0.06	20.2

✓ - overlap

✗ - no overlap

This table shows a summary of the results obtained from comparing COVER/COMPOSITE to GISTIC.

Downstream analyses

CNV-association analysis

One important use of the identified common CNV regions is for group comparisons in association studies. For each region we test whether certain CNVs are over-represented in one group compared to the others. Typically, the Fisher's exact test or chi-squared test for contingency tables can be used. The test can be carried out for all identified common CNV regions and the issue of multiple testing can be dealt with using the false discovery rate (FDR) assessment. (See [Additional file 1] on how to use the package for such analyses.)

As an illustration we performed an association analysis on the common regions identified in the 112 control subjects using the optimal parameters for COVER and COMPOSITE. The subjects were grouped by ethnicity (YRI, CHBJPT and CEU). Both methods showed that there were a number of highly-significant CNV regions with p-value < 1e-06. Two of these regions were detected by both methods. The first one is a 16.2 kb deletion in chromosome 2 (genomic positions 203,004,035 to 203,020,242). This region occurs exclusively in the Yoruba population (17/37) and overlaps with the BMPR2 gene that has been linked to primary pulmonary hypertension [22]. The second region is a 4.6 kb deletion in chromosome 4 (genomic positions 20,982,707 to 20,987,259) that occurs among Yoruban (19/37) and CHBJPT (4/29). This region overlaps with the KCNIP4 gene that is known to interact with presenilin, a protein that has been reported to be involved in early-onset Alzheimer's disease [23].

Principal component analysis of CNV profiles

We also perform principal component analyses (PCA) to obtain informative plots of population differentiation in the CNV profiles (see [Additional file 1] for more information). For the HapMap samples, the first two components obtained using the optimal COVER parameters separate the Yoruba population (YRI) from the Caucasian(CEU) and Asian(CHBJPT) populations, but the other two populations are not very well separated (Figure S1 in the [Additional file 1]). A better separation between the CEU and CHBJPT populations is achieved using the third and fourth components(see Figure 7(a)) and the separation is further improved when we use CLUSTER to refine the CNV regions identified by COVER (Figure 7(b)).

Conclusions

We have described and compared two different methods for identifying common CNV regions. Using 112 HapMap samples, we have shown that these methods produce common CNV regions that mostly follow Hardy-Weinberg equilibrium (HWE). For the eight HapMap samples where we compared the regions we identified to the reference CNV regions found by sequencing [10], the discordance rates can be as high as 80%, but this can be reduced to 60% by considering CNVs with higher confidence scores, thus showing the importance of

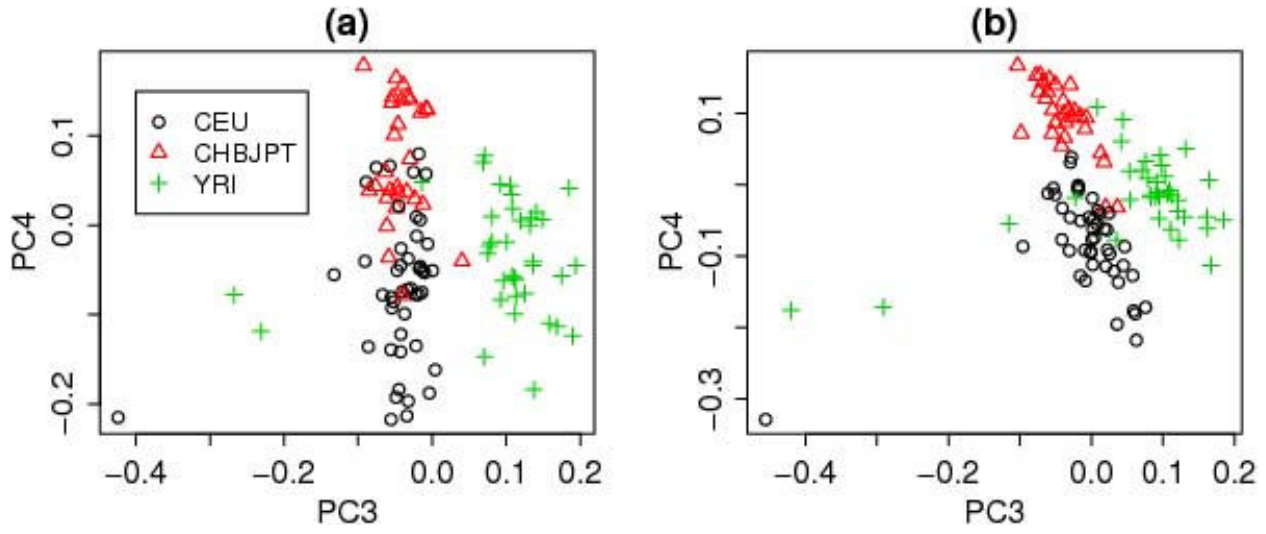


Figure 7 The third and fourth principal components. (a) Using COVER (with u = 3 and c = 60). (b) the same as (a) but using the output of complete-linkage CLUSTER (`with cluster.limit = 0.6`).

further processing of the CNVs. The high level of discordance itself might be due to an inherent limitation in the SNP array as the platform for CNV detection, but perhaps also due to imperfection in the sequencing-based results. Further works are needed to explain the discordance level.

When we compared our methods to previously published methods, STAC and GISTIC, we found that our methods are better at identifying low-frequency CNVs. Moreover, STAC is rather rigid and insensitive to the actual breakpoints of a CNV region, because if two consecutive windows are reported as significant, we do not know if there is one large CNV which spans both windows, or two separate and distinct CNVs. Although we can decrease the window size to increase the resolution, in practice, decreasing the window size beyond a certain point will incur too much computational burden. Another limitation of previous methods is the lack of consideration of individual-specific confidence scores. This means that all samples contribute equally to the calculation of the statistic used to identify the common regions, while in fact, there is bound to be inter-sample variability, where some CNVs are more likely to be true positives than others.

The results of COVER and COMPOSITE are similar in terms of discordance rates and HWE violation rates, but COMPOSITE appears to be better at identifying rare regions. The HWE violation rates are useful in determining the choice of parameter values for COVER and COMPOSITE. For this particular data set, we observed a steeper reduction in HWE violation rates when we used COVER with a confidence score threshold set above the median or higher. For COMPOSITE, a more noticeable reduction in HWE violation rates was observed when we set v to the 94th percentile. For a new dataset, we encourage users to choose the confidence score and COMPOSITE score parameter thresholds for which steeper reduction in HWE violation rates can be observed.

When using COVER, the minimum number of individuals inside a common region (u) needs to be specified as well. If we are interested in rare variants in addition to the common variants, then it makes sense to set u = 1. Otherwise, $u \geq 2$ should be used. A higher u will result in the identification of fewer, but more highly-recurrent CNV regions. In our experience with the HapMap samples, clustering results produce better separation of the ethnic groups than indicated by the initial common CNV regions. In comparison with the highly-validated CNVs from Conrad et al. [17], the concordance rate of COVER also improves after refinement with CLUSTER. So, in summary, we recommend users to further refine the identified common CNV regions using CLUSTER.

Additional file 1: The supplementary report documents details on how to use the R package cnvpack for the various analyses described in this paper.

Additional file 2: This table shows details of the regions found by COVER.

Additional file 3: This table shows details of the regions found by COMPOSITE.

Acknowledgements
This work was supported by a grant from the Swedish Research Council and National University of Singapore (NUS) Start-up Grant No. R-186-000-103-133.

Author details
[1]Department of Epidemiology and Public Health, National University of Singapore, 16 Medical Drive, 117597, Singapore. [2]Centre for Molecular Epidemiology, National University of Singapore, 30 Biopolis Street, 138671, Singapore. [3]Department of Biomedical Sciences and Biotechnology, University of Brescia, Viale Europa, 11 25123 Brescia, Italy. [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, Stockholm 17177, Sweden. [5]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, 28 Medical Drive, 117456, Singapore.

Authors' contributions
TSM and AS contributed equally to this work; TSM, AS, SC and YP conceived the study, performed data analysis and wrote the manuscript. KCS and CKS conceived the study. All authors read and approved the final manuscript.

Received: 15 October 2009 Accepted: 22 March 2010
Published: 22 March 2010

References

1. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5**:557-572.
2. Rueda OM, Diaz-Uriarte R: **Flexible and accurate detection of genomic copy-number changes from aCGH.** *PLoS Computational Biology* 2007, **3(6)**: e122.
3. Erdman C, Emerson JW: **A fast Bayesian change point analysis for the segmentation of microarray data.** *Bioinformatics* 2008, **24**:2143-2148.
4. Pique-Regi R, *et al*: **Sparse representation and Bayesian detection of genome copy number alterations from microarray data.** *Bioinformatics* 2008, **24**:309-3182.
5. Pique-Regi R, *et al*: **Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA.** *Bioinformatics* 2009, **25(10)**:1223-1230.
6. Wang K, *et al*: **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.** *Genome Research* 2007, **17**:1665-167.
7. Colella S, *et al*: **QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data.** *Nucleic Acids Research* 2007, **35**:2013-2025.
8. Rueda OM, Diaz-Uriarte R: **Finding Recurrent Regions of Copy Number Variation: A Review.** *Collection of Biostatistics Research Archive* 2008, Art42.
9. Diskin SJ, *et al*: **STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments.** *Genome Research* 2006, **16**:1149-1158.
10. Kidd JM, *et al*: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56-64.
11. Beroukhim R, *et al*: **Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma.** *PNAS* 2007, **104(50)**:20007-20012.
12. Van Wieringen WN, Wiel Van De MA, Ylstra B: **Weighted clustering of called array CGH data.** *Biostatistics* 2008, **9(3)**:484-500.
13. Eisen MB, *et al*: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863-14868.
14. Jong K, *et al*: **Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors.** *Oncogene* 2007, **26**:1499-1506.
15. Everitt BS, *et al*: *Cluster Analysis* Arnold, 4 2001.
16. McCarroll SA, *et al*: **Integrated detection and population-genetic analysis of SNPs and copy number variation.** *Nature Genetics* 2008, **40**:1166-1174.
17. Conrad DF, *et al*: **Origins and functional impact of copy number variation in the human genome.** *Nature* 2009.
18. Hupe P, *et al*: **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.** *Bioinformatics* 2004, **20(18)**:3413-3422.
19. Guttman M, *et al*: **Assessing the significance of conserved genomic abberations using high resolution genomic microarrays.** *PLoS Genetics* 2007, **3(8)**:e143.
20. Locke DP, *et al*: **Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome.** *American Journal of Human Genetics* 2006, **79**:275-290.
21. Redon R, *et al*: **Global variation in copy number in the human genome.** *Nature* 2006, **444**:444-454.
22. Lane KB, Consortium TIP, *et al*: **Heterozygous germline mutations in BMPR2, encoding a TGF-beta receptor, cause familial primary pulmonary hypertension.** *Nature Genetics* 2000, **26**:81-84.
23. Hutton M, Hardy J: **The presenilins and Alzheimer's disease.** *Human Molecular Genetics* 1997, **6**:1639-1646.

doi:10.1186/1471-2105-11-147
Cite this article as: Mei *et al.*: Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* 2010 **11**:147.

Supplementary information for

Identification of common copy-number variation (CNV) regions using high-density SNP array

Teo Shu Mei, Agus Salim, Stefano Calza, Ku Chee Seng, Chia Kee Seng, and Yudi Pawitan

The following R commands show how users can use the *cnvpack* package to identify common copy-number regions among 112 HapMap samples that are part of the Illumina iControl Database. The input files are PennCNV outputs based on Illumina 1M beadchip. We will only analyze common CNV regions (CNVR) detected in chromosome 1 to 22.

Download and install the *cnvpack* from http://www.meb.ki.se/~yudpaw

After installation, load the package in R using:

```
require(cnvpack)
```

The main functions in the package are summarized here:

1. `read.cnv` – reads in outputs from any CNV-detection software.

2. `setreg` – sets up common CNV regions using one of the three proposed methods, using user-specified parameters. The output will contain a list of common CNV regions and for each individual, a list of individual-specific common CNV regions and the discrete copy number calls.

3. `plot` – displays clusters of regions within each common region: only available when CLUSTER method is used.

4. `hwe.cnv` – tests Hardy-Weinberg Equilibrium (HWE) of diallelic CNV regions.

5. `summary` – prints descriptive statistics of size and frequencies of common CNV regions, conducts CNV-association analysis. The output is a list of common CNV regions with their frequency distribution and p-values from the association analysis, adjusted for multiple testing.

We read in the PennCNV output (`input="penncnv"`) which contains the list of detected individual CNVs for the 112 HapMap samples. This file contains confidence score values for the individual CNVs (`conf = TRUE`), as well as the list of genes within the individual CNVs (`annotated=TRUE`). The `read.cnv` function will then return a data frame of class `cnv`.

```
out<-
read.cnv(filename="112hapmap.txt",conf=TRUE,annotated=TRUE,
input="penncnv",sep="\t")
```

Note that if the output is from PennCNV or QuantiSNP, there must be a column indicating confidence scores.
Outputs from other software can also be read in, provided it contains the following information in tab delimited columns: (1) chr: chromosome where CNV region is located, (2) sample: sample name, (3) cn: the detected integer copy-number with cn=2 indicating normal copy number, (4) startsnp: SNP or copy-number probe where the region starts and (5) endsnp: SNP or copy-number probe where the region ends and (6) confidence (optional): a score that reflects how confident the CNV-detection algorithm is in calling the integer copy number. The magnitude of the score reflects the level of confidence. If confidence scores are not available, set `conf=FALSE`.

Next, we load the Illumina IM beadchip annotation file and the phenotype data for our samples. The annotation file contains a list with components “Name”, “Position", "Chromosome" and "Chr.num"; “Name” contains the name of the SNP or CNV marker, “Position” contains the position of the markers in the genome , “Chromosome” specifies the Chromosome the marker is in (1,2,3,…,22, X, Y, XY, (M)itochondrial) and “Chr.num” is the numeric form of “Chromosome”, where “X” is coded as “23”, “Y” is coded as “24”, “XY” is coded as “25” and “M” is coded as “26”. The phenotype data is a data frame with columns “sample”, “gender” and “ethnic” (other covariates may be added).

```
load('ann_illumina1M.Rdata')
```

```
head(ann$Name)
```

```
[1] rs12354060 rs2691310  rs2531266  rs4477212  rs4124251  rs8179466
1072820 Levels: cnv10000p1 cnv10003p1 cnv10003p5 cnv10004p1 ... SNP98
```

```
head(ann$Position)
```

```
[1]  10004  46844  59415  72017  97215 224176
```

```
head(ann$Chromosome)
```

```
[1] 1 1 1 1 1 1
26 Levels: 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 3 4 5 6 7 8 9 M X ... Y
```

```
head(ann$Chr.num)
```

```
[1] 1 1 1 1 1 1
```

Next, we load the data frame containing gender and ethnicity information of the 112 samples.

```
load ('112hapmap_pheno.Rdata')
head(pheno)
```

```
    sample gender ethnic
1 NA06985  FALSE    CEU
2 NA06991  FALSE    CEU
3 NA06994   TRUE    CEU
4 NA07000  FALSE    CEU
5 NA07029   TRUE    CEU
6 NA07345  FALSE    CEU
```

We now use the `setreg` function to set up the common CNV regions for these 112 individuals. The Cumulative Overlap Using Very Reliable Regions (COVER) method was used (`method="COVER")`. The minimum acceptable confidence score is set to be equal to the 60th percentile of all reported confidence scores `(high.conf=60)`. Inside a genomic location in a common CNV region, there must be at least 3 individuals whose individual CNVs overlap with that genomic location `(LIM=2), with` the deletion and duplication regions considered simultaneously `(cnv.abnormality= "both").` If confidence scores are not available, set `high.conf = NA.`

```
cnvr<-setreg(out,anno.list = ann,pheno.data=
pheno,high.conf = 60,LIM = 2,method = "COVER",
cnv.abnormality = "both")
```

The `cnvr$reg$loc` object containing the list of all identified common CNV regions with their positions in term of probes number:

```
 head(cnvr$reg$loc)
$`limit=2`
         st      en st.chr regmin
1         5      47      1    -15
2       205     248       1     -7
3       269     283       1     -4
4      2681    2695       1     -4
5      4201    4208       1     -3
6      4367    4381       1     -5
7      5380    5392       1     -4
8      5395    5399       1     -3
.
.
.
443 1029556 1029574     22     -8
```

The `cnvr$reg$regID` is a list. For each individual, it contains the CNVR ID (corresponding the ID in `cnvr$reg$loc)` for which that individual has copy-number variation. The corresponding integer copy-number call is given in `cnvr$reg$call.` The ordering of the individual in `cnvr$reg$regID` corresponds to the way they are ordered in `cnvr$pheno,` so that `cnvr$reg$regID[[1]]` contains the CNVR for individual on the first row of `cnvr$pheno; cnvr$reg$regID[[2]]` contains the CNVR for individual on the second row of `cnvr$pheno` and so on.

Test for Hardy-Weinberg Equilibrium

We can test for Hardy-Weinberg Equilibrium of common diallelic CNV regions when `cnv.abnormality= "both"` in `setreg.`

```
cnvr.hwe = hwe.cnv(cnvr)
```

The main component of interest in `cnvr.hwe` would be `cnvr.hwe$group1$fdr` which gives the pvalue from testing HWE on each CNVR in `cnvr$reg$loc`, corrected

for multiple testing using Benjamini-Hochberg method (the pvalue is set to NA for multiallelic CNVR).

```
head(cnvr.hwe$group1$fdr)
[1] 0.2992617 0.6346215 0.7874008 0.8400692 0.8400692
0.7874008
```

The ordering of the adjusted pvalues in `cnvr.hwe$group1$fdr` corresponds to the CNVR in `cnvr.hwe$group1$reg,` which in turns corresponds to `cnvr$reg$loc.`

```
head(cnvr.hwe$group1$reg)
[1] 1 2 3 4 5 6
```

Downstream Analyses

Comparing between-group CNV frequency

We can obtain a summary of the number of CNVs and distribution of CNV length for each level of `group` variable. Using `test= TRUE`, we can compare between-group CNV frequency. A output of the list of common CNVs with their profiles and frequency distribution as well as p-values and false discovery rates will be saved in LIST of CNV regions.xls (see Supplementary Table I) .Here, we are comparing between ethnicity, where the ethnicity information was located in the third column of our phenotype data (`group= 3`).

```
summary(cnvr,anno.list= ann,test= TRUE,group= 3)
```

A group-by-group summary of the CNV regions will be displayed as follows:

```
 CNV Region group-by-group summary:

 group=CEU
Number of unique CNV regions across individuals = 322
Average length of CNV regions 71039.51
Median length of CNV regions 19175
Distribution of length

    (0,1e+03] (1e+03,1e+04] (1e+04,5e+04] (5e+04,1e+06]
            6            90           148            73

 CNV Region group-by-group summary:
 group=CHJP
Number of unique CNV regions across individuals = 175
Average length of CNV regions 106601.0
Median length of CNV regions 23551
Distribution of length

    (0,1e+03] (1e+03,1e+04] (1e+04,5e+04] (5e+04,1e+06]
            5            56            54            55

 CNV Region group-by-group summary:

 group=YRI
Number of unique CNV regions across individuals = 362
```

```
Average length of CNV regions 63749.46
Median length of CNV regions 18185
Distribution of length

    (0,1e+03] (1e+03,1e+04] (1e+04,5e+04] (5e+04,1e+06]
            5           102           176            74
Correlation of CNV Regions Frequency
           group=CEU group=CHJP group=YRI
group=CEU  1.0000000  0.5093919 0.1071928
group=CHJP 0.5093919  1.0000000 0.2491267
group=YRI  0.1071928  0.2491267 1.0000000
```

Principal Component Analysis (PCA) of CNV profiles

To perform PCA on the CNV profiles, we need to create A, a $n \times p$ matrix of absence-presence, where n is the number of individuals and p is the number of CNV regions. The j^{th} element of the i^{th} row will be equal to 1 if the i^{th} subject has CNV in the j^{th} region and 0 otherwise. We can extract this information from the `regID` component of `reg` list, an output of `setreg.princomp(t(x))` will perform the analysis on $\frac{1}{p}AA^T$. A matrix of scatterplots of the first five principal component loadings is shown in Figure S1. For a clearer picture, we also plot loading 3 versus loading 2 in Figure S2. We can see that the three ethnic groups are well separated. A summary of the proportion of data explained by the different components can be obtained by using `summary ( )`. There are in total 112 principal component loadings, and the first 5 components explain about 27% of the variance in the data.

```
# The following commands perform PCA on integer copy-number
n = 112 #number of individuals
ncol = max(unlist(cnvr$reg$regID)) #get the largest CNVR ID
reg = unique(unlist(cnvr$reg$regID)) #get a vector of
unique CNV region ID
x = matrix(2,nrow=n,ncol=ncol)
for(i in 1:n){   #for each individual
       x[i,unlist(cnvr$reg$regID[[i]])] =
unlist(cnvr$reg$call[[i]])}

out = princomp(t(x))

#Figure S1
pairs(out$load[,1:5],main="First five principal component
loadings",
col=as.numeric(factor(cnvr$pheno$ethnic)),pch=substring(cnv
r$pheno[,3],3,3),cex = 0.6)
```

```
summary(out)
```

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     0.8847292 0.66110492 0.58680472 0.44996690 0.40621197
Proportion of Variance 0.1160093 0.06477583 0.05103398 0.03000775 0.02445556
Cumulative Proportion  0.1160093 0.18078518 0.23181916 0.26182690 0.28628247
                          Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.39876638 0.39198306 0.37636555 0.37189963 0.36877064
Proportion of Variance 0.02356727 0.02277230 0.02099384 0.02049858 0.02015510
Cumulative Proportion  0.30984974 0.33262203 0.35361588 0.37411446 0.39426955
.
.
.
                         Comp.11    Comp.12    Comp.13    Comp.14    Comp.15
                          Comp.108     Comp.109     Comp.110     Comp.111
Standard deviation     0.0758989870 0.0740548526 0.0701807397 0.0681210169
Proportion of Variance 0.0008537765 0.0008127917 0.0007299752 0.0006877561
Cumulative Proportion  0.9972214494 0.9980342411 0.9987642163 0.9994519725
                          Comp.112
Standard deviation     0.0608086094
Proportion of Variance 0.0005480275
Cumulative Proportion  1.0000000000
```

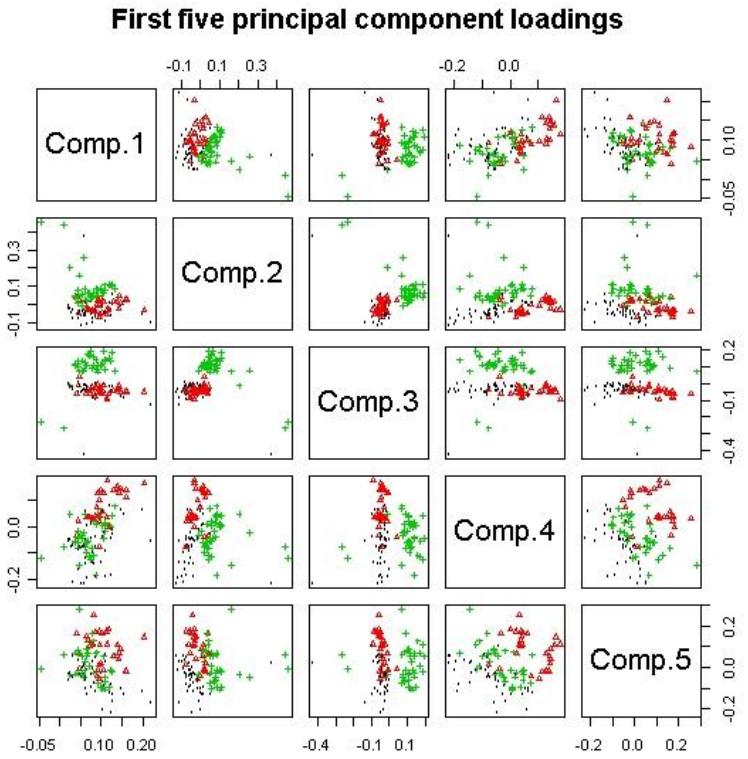


Figure S1. The first five principal component loadings based on the CNV profiles of the European population CEU (black), the Yoruba population (green) and the Asian population (red) based on CNV region profiles identified using COVER with u=3 and c=60.

COVER with CLUSTER

We refine the CNV regions identified by COVER above and look for clustering of individual regions (see Figure S3 for evidence of clustering). Here, we use hierarchical clustering with complete linkage and require that within-cluster similarity to be at least 60% (see Table S1 for effects on linkages and cut-off on number of clusters).

```
cnvr <-setreg(out,anno.list= ann,pheno.data=
pheno,high.conf= 60,LIM= 2,method= "COVER",
cnv.abnormality=
"both",cluster.method='complete',cluster.limit=0.6)
```

We perform principal component analysis (PCA) based on the common CNV region profiles identified by this configuration of threshold parameters.

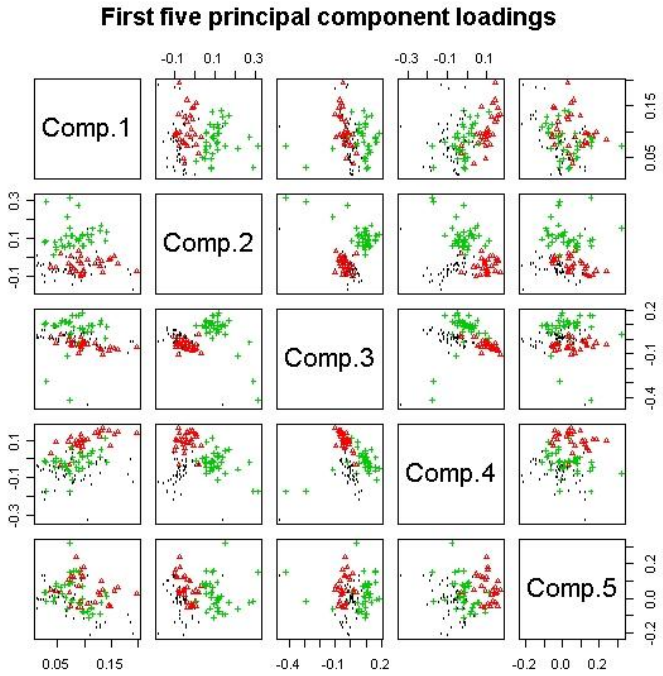


Figure S2. The first five principal component loadings based on the CNV profiles of the European population CEU (black), the Yoruba population (green) and the Asian population (red) based on CNV region profiles identified using COVER with u=3 and c=60 and refined using CLUST with complete linkage and cluster.limit=0.6

```
## PCA
p = dim(cnvr$reg$loc[[1]])[1]
n = length(cnvr$reg$regID)    #no. of individuals
x = matrix (2, nrow = n, ncol = p)
for(i in 1:n){    #for each individual
```

```
index =
match(as.numeric(unlist(cnvr$reg$regID[[i]])),as.numeric(rownames
(cnvr$reg$loc[[1]])))
x[i,sort(index) ] = unlist(cnvr$reg$call[[i]])
}
out  = princomp(t(x))
```

Figure S2

```
pairs(out$load[,1:5],main="First five principal component
loadings",
col=eth,pch=eth,cex = 0.6)
```

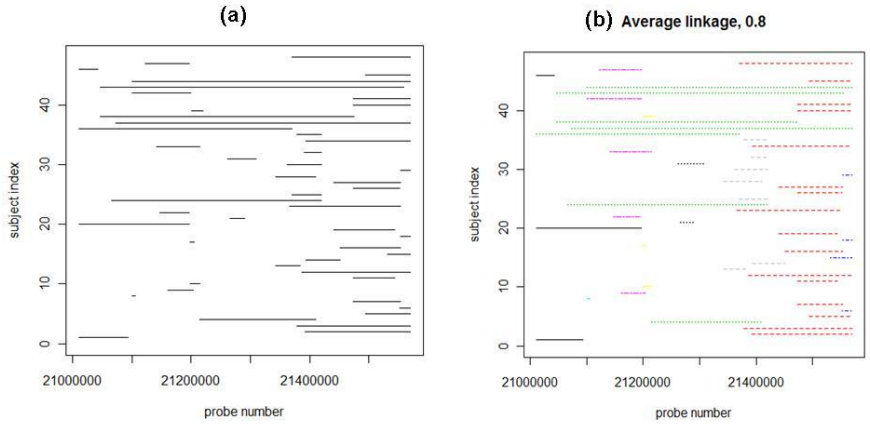


Figure S3. (a) A common CNV region identified using COVER with u=3 and c=60, exhibiting clusters of individual regions and (b) identified clusters of individual regions obtained using average linkage and 80% similarity cut-off.

Dissimilarity cut-off	Number of Clusters		
	Single Linkage	Average Linkage	Complete Linkage
0.01	1	2	9
0.05	2	5	9
0.10	2	6	10
0.20	2	9	12
0.30	4	12	14
0.50	11	17	21
0.60	16	21	23
0.70	24	27	28

Table S1. The number of clusters identified in CNV region in Figure S3 as a function of dissimilarity (1-similarity) cut-off point and linkage.

BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 11 2011, pages 1555–1561
doi:10.1093/bioinformatics/btr162

Data and text mining

Advance Access publication April 5, 2011

Multi-platform segmentation for joint detection of copy number variants

Shu Mei Teo[1,2,3], Yudi Pawitan[2], Vikrant Kumar[4], Anbupalam Thalamuthu[4], Mark Seielstad[4], Kee Seng Chia[1] and Agus Salim[1,*]

[1]Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, National University of Singapore, Singapore, [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore and [4]Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: With the expansion of whole-genome studies, there is rapid evolution of genotyping platforms. This leads to practical issues such as upgrading of genotyping equipment which often results in research groups having data from different platforms for the same samples. While having more data can potentially yield more accurate copy-number estimates, combining such data is not straightforward as different platforms show different degrees of attenuation of the true copy-number or different noise characteristics and marker panels. Currently, there is still a relative lack of procedures for combining information from different platforms.

Results: We develop a method, called MPSS, based on a correlated random-effect model for the unobserved patterns and extend the robust smooth segmentation approach to the multiple-platform scenario. We also propose an objective criterion for discrete segmentation required for downstream analyses. For each identified segment, the software reports a *P*-value to indicate the likelihood of the segment being a true CNV. From the analyses of real and simulated data, we show that MPSS has better operating characteristics when compared to single-platform methods, and have substantially higher sensitivity compared to an existing multiplatform method.

Availability: The methods are implemented in an R package MPSS, and the source is available from http://www.meb.ki.se/~yudpaw.

Contact: agus_salim@nuhs.edu.sg

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on December 17, 2010; revised on March 8, 2011; accepted on March 26, 2011

1 INTRODUCTION

Copy-number variants (CNVs) are defined as duplications or deletions in the number of copies of a DNA segment (larger than 1 kb in length) when compared to a reference genome. Currently, common technologies used to detect CNVs include high-density single nucleotide polymorphism (SNP) arrays and comparative-genomic hybridization (CGH) arrays. In recent years, whole-genome studies using commercial genotyping arrays to detect CNVs have been rapidly expanding. With decreasing cost of commercially available platforms and the fast evolution of these platforms, it is not unusual for research groups to have data from multiple platforms for each sample. For example, The Cancer Genome Atlas Research Network, a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to explore genomic changes involved in human cancers, used Agilent 244K, Affymetrix SNP 6.0 and Illumina 550K platforms to measure copy number alterations in its pilot study. Our own collaborators, and perhaps many other researchers, collected genotype data using both Illumina HumanHap300 and HumanHap240S arrays for each sample in order to get higher genome coverage.

Marker density is an important factor for comprehensive and accurate detection of CNVs and their breakpoints, and different platforms have different probe coverage and density; see Curtis *et al.*(2009) for a summary of probe coverage of the different platforms in the different chromosomes. Combining data from different platforms can potentially yield more precise and accurate detection of CNVs and its breakpoints. However, combining such data is not straightforward because it is known that estimates from different platforms show different degrees of attenuation of the true copy-number changes (Bengtsson *et al.*, 2009) as well as different noise characteristics. Furthermore, different platforms have different marker panels and molecular assay methods (Zhang *et al.*, 2010). Currently, there is still a relative lack of formal procedures for combining information from different platforms for copy-number calling. Most studies with multiple platforms interrogating the same samples process the data from the different platforms independently, then combine the segments in an *ad hoc* manner. This approach does not fully utilize information from the different platforms, and when the segmented results from the different platforms differ, it is difficult for researchers to come to a consensus in a statistically rigorous manner.

One published method, multiple platform circular binary segmentation (MPCBS) (Zhang *et al.*, 2010), is able to jointly use information from different platforms for CNV calling. The MPCBS method extends the circular binary segmentation (CBS) algorithm (Olshen *et al.*, 2004) by detecting coupled changes in multiple sequences. Briefly, it uses a weighted sum of *t*-statistics

*To whom correspondence should be addressed.

from a generalized log-likelihood ratio of a multiplatform model and pools statistical evidence across platforms during segmentation.

The proposed multiplatform smooth segmentation (MPSS) method extends Huang *et al.* (2007)'s smoothseg algorithm, which is based on the Cauchy random-effect model that allows jumps in the underlying copy-number patterns to the multiple platforms scenario. The algorithm computes the estimated random-effect estimates that capture the underlying copy-number patterns, and is applicable to both germ-line and tumor DNA as long as the data has been appropriately normalized. As we are often interested in discrete segments of deletions, normal copies and duplications for downstream analysis such as CNV association studies, we also develop an objective method to obtain the discrete segmentation. From analyses of real and simulated data, MPSS performs well compared to single-platform methods, and shows substantially higher sensitivity compared with MPCBS.

2 METHODS

We first describe the correlated random-effect model for the unobserved pattern. For each individual, denote $X \equiv \{x_1,...,x_n\}$ as the union of the genomic locations of probes from the different platforms, with $x_1 < x_2 < ... < x_n$. Denote $Y_j \equiv \{y_{x_{1j}},...,y_{x_{nj}}\}$ as the set of $\log_2$-intensity ratios from platform j, n_j is the number of probes in platform j. Let $N = \sum n_j$. We consider the model:

$$y_{x_{ij}} = f_j(x_{ij}) + e_{x_{ij}} \tag{1}$$

where f_j is the unknown platform-specific random-effects; the platform-specific errors are independent and identically t-distributed with location parameter 0, unknown dispersion parameter σ_j and k degrees of freedom. We assume the errors and the random-effects to be independent. The error structure was chosen to be t-distributed to incorporate a heavy-tailed structure that can deal with outliers in the observations. We simplify (1) to

$$y_{x_{ij}} = f(x_{ij}) + e_{x_{ij}} \tag{2}$$

such that $f(.)$ is a random effect parameter common to all platforms. This simplification is justified when data from the different platforms are well normalized, because the different platforms are measuring the same underlying copy-number pattern. If not, a normalization procedure has to be applied first. Note that the error term is still platform-specific. In matrix form, we write (2) as

$$Y \equiv Zf + \varepsilon$$

where Z is the model matrix determined by the observed x's and the choice of basis functions. We use the observed x's as knots and choose the zero-order B-splines. Hence, Z is the N by n model design matrix that indicates the genomic locations of the probes from the different platforms, meaning that the row of Z associated with the original data y_{ij} has value one at the i-th location and zero otherwise. The smoothness of f can be expressed by assuming that the scaled second-order differences $a_i^* \equiv \frac{\Delta^2 f_i}{(\Delta x_i)^2}$ are i.i.d. with some distribution. Since f is mostly smooth, the size of $a_i \equiv \Delta^2 f_i = (\Delta x_i)^2 a_i^*$ is very small relative to the local noise. So, there will be little difference whether we specify the model on a_i^* or a_i. For convenience, we shall use the latter. We choose the Cauchy distribution with location 0 and scale factor σ_f^2. The Cauchy distribution has been used to deal with jumps in the underlying patterns, with desirable results (see Huang *et al.*, 2007, 2009).

2.1 Estimation of f via maximum likelihood

We derive an iterative weighted least squares algorithm by maximizing the likelihood of the Cauchy random-effects model (see Huang *et al.*, 2007 and Pawitan, 2001, pp. 464–466). The log-likelihood based on y and f, assuming σ_j^2's and the smoothing parameter $\lambda = \frac{ave(\sigma_j^2)}{\sigma_f^2}$ are known, is $\log L(f, \sigma_f^2, \sigma_j^2) = \log p(y|f) + \log p(f)$. The first term comes from the t-density with k degrees of freedom: For all (i,j) where platform j has a probe at location i,

$$\log p(y_{ij}|f) = c - \frac{1}{2}\log(\sigma_j^2) - \frac{k+1}{2}\log\left\{k + \frac{(y_{ij}-f_i)^2}{\sigma_j^2}\right\} \tag{3}$$

where c is a constant. The second term comes from the Cauchy model with location 0 and scale factor σ_f:

$$\log p(f) \equiv l(a) = -(n-2)\log(\pi\sigma_f) - \sum_{i=1}^{n-2}\log\left(1 + \frac{a_i^2}{\sigma_f^2}\right) \tag{4}$$

Differentiating (3) with respect to f, we get:

$$\frac{\partial \log p(y|f)}{\partial f} = Z'WY - Z'WZf \tag{5}$$

where W is a N by N diagonal matrix with diagonal elements $w_{ij} = \frac{k+1}{k\sigma_j^2 + (y_{ij}-f_i)^2}$, associated with the corresponding original data y_{ij}. In scalar form, the i-th element of (5) can be written as $\sum_j w_{ij}(y_{ij} - f_i)$. Differentiating (4), we obtain:

$$l'(a) = -D^{-1}a \tag{6}$$

where $a = \Delta^2 f$, denoted by Δ^2 the $(n-2)$ by n matrix that represents the second-order difference operator and

$$D^{-1} = \text{diag}\,[2/(\sigma_f^2 + a_i^2)]$$

Combining (5) and (6), we obtain the score function, the first derivative of $\log L(f, \sigma_f^2, \sigma_j^2)$ as:

$$S(f) = (Z'WY - Z'WZf) - (\Delta^2)'D^{-1}(\Delta^2)f$$

Setting $S(f) = 0$, we get

$$[Z'WZ + (\Delta^2)'D^{-1}(\Delta^2)]f = Z'WY \tag{7}$$

We estimate f from (7) by exploiting the band-limited property of $[Z'WZ + (\Delta^2)'D^{-1}(\Delta^2)]$ and use well-tested fortran subroutines available in Linpack (see Huang *et al.*, 2007 and Dongarra *et al.*, 1979).

2.2 Estimation of σ_j

Given $\hat{f}$, at each probe position i, the deviance is defined as

$$d_i = (k+1)\log\left\{1 + \frac{(y_{ij} - \hat{f}_i)^2}{k}\right\}$$

This can be approximated by the gamma distribution with mean μ_i and dispersion ϕ. To estimate μ_i, we use a generalized linear model with a log-link function, so $h(\mu) = \log(\mu)$ and $h(\mu_i) = x_i^t\alpha$, where the dimension of x_i and α is equal to the number of platforms. We solve using IWLS with robust weights:

(1) Start with an initial α_0. We estimate ϕ once using $\hat{\phi} = \frac{var(d_i)}{\bar{d_i}^2}$.

(2) We write

$$Y^* = X\alpha + e^* \tag{8}$$

where Y^* is called the working vector with elements

$$y_i^* = \frac{\partial h}{\partial \mu}(d_i - \mu_i^0) + x_i^t\alpha_0 \tag{9}$$

$e_i^* = \frac{\partial h}{\partial \mu}e_i$ and $\text{var}(e_i^*) = (\frac{\partial h}{\partial \mu})^2\phi\mu_i^2 = \phi$

(3) We use robust weights

$$w_i = \frac{1}{\text{var}\,(e_i^*)} \times w_{\text{huber}}, \tag{10}$$

where w_{huber} is the commonly used Huber weight function defined as

$$w_{\text{huber}}(e^*) = \begin{cases} 1 & \text{if } |e^*| <= c_j \\ c_j/|e^*| & \text{if } |e^*| > c_j \end{cases}$$

where $c_j = 1.345\sigma_j$. As an initial estimate, we use a robust measure of spread, $\hat{\sigma_j} = \text{median}(|e_j^*|)/0.6745$. Then α can be solved using the usual weight least squares solution:

$$\hat{\alpha} = (X'WX)^{-1}X'WY^* \tag{11}$$

(4) We iterate between steps (2) and (3) until convergence. Then we obtain $\hat{\sigma_j^2} = e^{\hat{\alpha}_j}$. In subsequent sections, if there is a need for a single σ estimate, we use the average of the σ_js.

2.3 Choosing optimal λ

The degrees of freedom associated with f is given by (Pawitan, 2001, p. 448)

$$df = \text{trace}\{(Z'WZ + (\Delta^2)'D^{-1}(\Delta^2))^{-1}Z'WZ\}$$

where W and D are computed using $\hat{f}$. This expression is hard to obtain computationally, so we use an approximation (Pawitan, 1996)

$$df \approx \sum_{k=1}^{n-2} \frac{\bar{w}}{\bar{w} + v_k^2/\bar{d}}$$

where $\bar{w}$ and $\bar{d}$ are the average diagonals of $Z'WZ$ and D, and

$$v_k = 2[1 - cos\{\pi(k-1)/n\}]$$

is the j-th eigenvalue of the second derivative matrix Δ^2. We choose λ that minimizes the Akaike information criterion (AIC)

$$\text{AIC}(\lambda) = -2\sum \log p(y_{ij}|\hat{f}) + 2df$$

2.4 Summary of MPSS algorithm

For a given $\lambda = \frac{\text{ave}(\sigma_j^2)}{\sigma_f^2}$, we employ the following algorithm:

(a) Start with an initial value for f_0 and σ_j^2's.

(b) Compute $\sigma_f^2 = \frac{\text{ave}(\sigma_j^2)}{\lambda}$.

(c) Compute $Z'WZ$ and D^{-1} and update f using (7).

(d) Update σ_j^2's as described in Section 2.2 .

(e) Repeat (b)–(d) until convergence.

2.5 *P*-values for segments

The Fisher information for f is the negative of the second derivative of the log-likelihood.

$$I(f) = Z'WZ + (\Delta^2)'D^{-1}(\Delta^2)$$

At convergence, the estimated variance matrix is

$$V = I^{-1}(f)$$

If we have a segment $\mathcal{S}$, defined for instance by setting a threshold, then $f_{\mathcal{S}}$ is a vector which contains the estimated values in the segment and zero everywhere else,

$$f_{\mathcal{S},i} = \begin{cases} \hat{f}_i \text{ if } i \text{ is in } \mathcal{S} \\ 0 \text{ otherwise} \end{cases}$$

The significance of the segment can be assessed using the statistic

$$\chi^2 = f'_{\mathcal{S}} V^{-1} f_{\mathcal{S}} \qquad (12)$$

To compute (12) without explicitly obtaining the inverse of a matrix with extremely large dimension, we write (12) as

$$f'_{\mathcal{S}}[Z'WZ + (\Delta^2)'D^{-1}(\Delta^2)]f_{\mathcal{S}} = \sum(f_{\mathcal{S},i}^2 w_i) + \sum(a_{\mathcal{S},i}^2 d_i),$$

where $a_{\mathcal{S},i}$ contains the second-order differences of $f_{\mathcal{S}}$ and d contains the diagonal elements of D^{-1}. We compare this statistic to the chi-squared distribution with q degrees of freedom, where q is the number of probes in $\mathcal{S}$. To adjust for multiple hypothesis testing involving a large number of segments, we compute the false discovery rate (FDR) for each segment.

2.6 Objective threshold segmentation

A segment whose random-effects parameter f consistently and significantly deviates from zero is evident of a deletion/duplication. We obtain potential copy-number segments by setting thresholds for $\hat{f}$, where duplications are sets of consecutive probes for which $\hat{f}$ is consistently greater than or equal to a specified threshold, and deletions are sets of consecutive probes for which $\hat{f}$ is consistently smaller than or equal to a specified threshold. For automatic threshold selection, users can pick the threshold that maximizes the total χ^2 values (scaled by its associated degrees of freedom).

To avoid oversegmentation, we merge the segments if the distance between adjacent segments is less than 5 kb. For each segment, we compute its associated *P*-value/FDR as described in the previous section. We further filter the segments by its length (those that are less than 1 kb are omitted), FDR and number of probes (minimum number of probes within segment is 10). A segment will also be omitted if the adjacent distance between 2 consecutive probes is larger than 100 times the median interprobe distance. All filtering parameters can be changed by the user. Users can also filter the segments by probe density (Number of probes/length of segment).

2.7 Removal of discrepant segments

For each segment identified by the MPSS algorithm, we test if the mean intensity from the different platforms differ using a t-test (corrected for autocorrelation, assuming the data has a first-order autoregressive structure) if there are two platforms and ANOVA if there is more than two platforms. We remove the segments where the FDR for the test is <0.01. We call these segments 'discrepant segments'. Discrepant segments are removed because the multiplatform algorithm assumes the signals from the different platforms are consistent with each other, hence signals from 'discrepant' segments are likely to be unreliable.

2.8 Comparisons using simulated data

We conduct a simulation study to evaluate the performance of MPSS as well as to compare against the MPCBS method. To get a realistic noise pattern, we use the empirical CNV profile of chromosome 1 of the Hapmap sample NA10851. We use data from both Affymetrix 6.0 and Illumina 1M platforms (see Section 3.1) and apply the MPSS algorithm with segmentation threshold of 0.05 and FDR threshold of 10^{-5}. We remove segments with <4 probes as extremely short segments are more likely to be false positives due to noise. We label the different segments of the chromosome as CNV or 'NULL'. In total, there are 12 CNV segments and 13 'NULL' segments. We perform the simulation study at three different noise levels; the input values are the smoothed intensities plus 0.5, 1, and 2 times the residuals from the respective platforms. Note that the smoothed intensities plus 1 times the residuals is exactly the original input intensities. We sample the 25 segments randomly with replacement and use their corresponding intensity values as input to the MPSS and MPCBS algorithms. We calculate the percentage of CNV probes that were correctly identified (sensitivity) and the percentage of 'NULL' probes that were correctly identified (specificity). We repeat the process 100 times by bootstrapping from the residuals.

Labeling the CNV segments using segments originally identified by the MPSS method may bias the analysis in favor of MPSS. Hence, we also repeat the whole process, labeling the CNV segments using segments identified by MPCBS, with a segmentation threshold of 0.05. After removing those with less than 4 probes, we are left with 6 CNV segments and 7 'NULL' segments.

2.9 Comparisons using real data

We compare MPSS against the single-platform smoothseg as well as MPCBS in a real data setting. We use the integer copy-numbers for a total of 5037 CNV loci from Conrad *et al.*(2010)'s study as well as McCarrol *et al.*(2008)'s study as a reference list. A set of 20 NimbleGen arrays, each comprising 2.1-million long oligonucleotide probes were used to first generate a new map of CNV locations. Subsequently, a customized Agilent CGH-platform

comprising of 105 000 long oligonucleotide probes was used to detect the loci and the genotypes were estimated for 450 HapMap samples using a Bayesian algorithm with stringent selection for optimal normalization and cluster locations for every locus [See Supplementary Material in Conrad *et al.* (2010) for more details]. We remove segments in the reference if the number of probes from the combined probe list from the two platforms we are using is less than 10 or if the segment size is less than 1 kb. There is a median of 163 CNV segments per individual.

It should be noted, however, that this reference list cannot really be considered the gold standard, as even sequencing data do not have 100% sensitivity and specificity in CNV detection (Xie *et al.*, 2009). For each method, we perform individual-specific comparisons with Conrad's CNVs and compute the number of bases that are called as CNV both by the method and by Conrad *et al.* We report the number of overlapping bases as a proportion to the total length of CNVs identified by the method and as a proportion of total length of Conrad's CNVs. While these may not be considered a 'true discovery rate' and 'sensitivity´, since Conrad's CNVs are well-validated, a higher proportion of overlap is an indication of better performance.

2.10 Implementation and computing time

The methods are implemented in an R package MPSS. The main inputs are vectors of genomic positions, chromosome numbers and $\log_2$-intensity ratios from each platform. It is recommended that users check if data from the different platforms are well-normalized. If not, background correction should be performed first; the package rsmooth from http://www.meb.ki.se/∼yudpaw can be used for background correction. All computations for this article was done on a 3 GHz Intel Core 2 Duo processor. For 1 individual, with more than 2.5 million combined probes from Affymetrix 6.0 and Illumina 1M, and for a user-specified λ, the algorithm takes <1 min. It takes <6 min if the AIC criteria is used to find the optimal λ.

3 RESULTS

3.1 Datasets and background correction

We use nine HapMap samples (International HapMap Consortium, 2005 and see Supplementary Materials for sample ID and population). These samples were previously genotyped by two SNP arrays (Illumina 1M and Affymetrix 6.0) in our research lab. We first perform background correction on the $\log_2$-intensity ratios from each platform using a robust smoother in the rsmooth package from http://www.meb.ki.se/∼yudpaw. This normalization assumes that the majority of the genome does not contain CNVs, which is the case for germline samples.

To investigate if the input intensities are well-normalized, we randomly sample a non-CNV segment of 100 consecutive probes and test if the mean intensity for the Affymetrix platform is equal to the mean intensity of the Illumina platform (using *t*-test corrected for autocorrelation). We repeat the process 1000 times and record the percentage of *P*-values that are less than 0.01. The normalization results look reasonable for all individuals with the percentage of *P*-values less than 0.01 ranging from 0.0043 to 0.02.

3.2 Estimated parameters

For each chromosome, we use the λ that minimizes the AIC criterion. A large variation in the optimal λ is observed across the genome, indicating the need for the selection of different λs for different chromosomes. For example, for individual NA19139, the optimal λ ranges from about 47 for Chromosome 15 to about 4900 for Chromosome 19; see Figure 1.

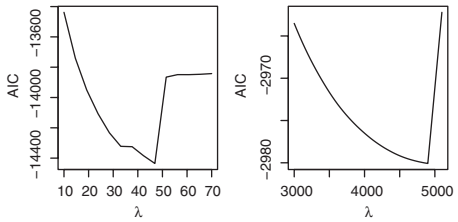


Fig. 1. AIC as a function of λ for data from chromosome 15 and 19 for individual NA19139.

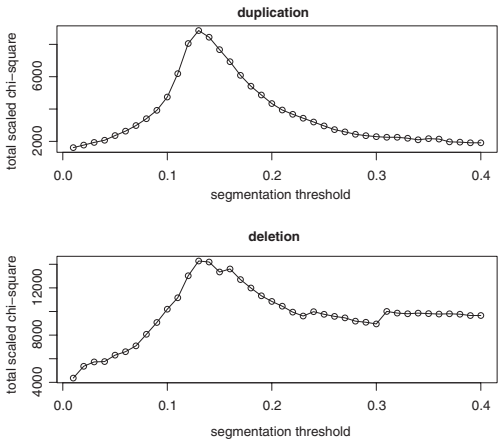


Fig. 2. Scaled total χ^2 as a function of segmentation thresholds (in absolute values) for individual NA19139.

3.3 Choice of thresholds

For each individual, we choose the deletion and duplication thresholds that give the largest total scaled chi-squared value. For individual NA19139, the deletion and duplication thresholds was chosen to be 0.13 (Fig. 2). At the chosen threshold values, after removing the discrepant segments (see Section 2.7), the algorithm identifies a median of 137, 129, 117 and 110 segments that passed the FDR threshold of 10^{-6}, 10^{-7}, 10^{-8} and 10^{-9}, respectively. The median length of the segments are 15.6, 16.1, 16.7 and 17.3 respectively.

At the same segmentation and FDR thresholds, the single platform algorithm identifies a median of 68, 63, 58 and 56 segments (median length of 39.6, 41.4, 43.3 and 42.8 kb) for the Illumina platform and 81, 77, 69 and 66 segments (median length of 40.8, 41.6, 44.5 and 45.1kb) for the Affymetrix platform.

We apply the MPCBS method on the signals, post-background correction, and use the modified Bayesian information criterion (BIC) approach as suggested by the authors to estimate the number of segments. The maximum number of change points per chromosome is set to 30. MPCBS outputs the breakpoints of the segments as well as the estimated response from each platform. We calculate the estimated response for each segment as the average of the responses from the two platforms. For each individual, we vary the segmentation thresholds such that the total length of CNVs identified by MPCBS is similar to MPSS. Similar results are obtained if we control the number of CNVs, so the results are not shown here. We require that the segments have a minimum of 10 SNPs, a minimum

Table 1. Sensitivity and specificity of MPSS and MPCBS using simulation data

	Sensitivity		Specificity	
Data input	MPSS	MPCBS	MPSS	MPCBS
MPSS segments				
1*residuals	0.779	0.098	>0.999	> 0.999
0.5*residuals	0.804	0.281	0.918	> 0.999
2*residuals	0.127	0.025	> 0.999	> 0.999
MPCBS segments				
1*residuals	0.527	0.192	0.999	>0.999
0.5*residuals	0.627	0.414	0.914	>0.999
2*residuals	0.177	0.098	>0.999	0.989

length of 1 kb and a maximum length of 1.1 Mb. The median length of the segments at these thresholds are 21.8, 21.8, 21.7 and 21.7 kb, respectively.

We also ran the single platform CBS algorithm. With $\alpha = 0.01$ and segmentation threshold of $\pm$ 0.01, 0.02, 0.03 and 0.04, we obtain a median of 30, 28, 26 and 26 segments (median length of 21.9, 19, 16.1 and 14.8 kb) for the Illumina platform and a median of 75, 73, 71 and 69 segments (median length of 37.8, 34.8, 34.3 and 31.5 kb) for the Affymetrix platform.

3.4 Comparison: simulated data

The average sensitivity and specificity across 100 bootstrap samples are summarized in Table 1. For most scenarios, both MPSS and MPCBS have high specificity (greater than 99%), though MPSS has slightly lower specificity when noise level is decreased. For both algorithms, sensitivity increases with decreased noise level and vice versa. In all cases, MPSS has substantially higher sensitivity than MPCBS. Mean sensitivity for MPSS can be as high as 80% when the noise level is decreased, whereas MPCBS only attains a mean sensitivity of about 41%. When noise level is high, both algorithms perform poorly—MPSS with a mean sensitivity of about 18% and MPCBS with a mean sensitivity of about 10%. However, note that with twice the magnitude of the residuals, the platform variability is increased to four times the original variability. With such high level of noise, unless the CNV signal is very strong, no algorithm is likely to identify the CNV.

3.5 Comparison: real data

When signals from the different platforms are consistent, we get increased power to detect the CNVs when we combine the information from the different platforms, especially in areas where a single platform has low density of probes. Figure 3a shows that the Illumina platform has a single probe in the deletion region, and while this probe exhibits strong evidence of a decreased intensity ($\log_2$-intensity ratio less than -3), the single platform approach was unable to identify the deletion. On the other hand, the Affymetrix platform has several probes in the region with moderately decreased $\log_2$-intensity ratio values, and the single platform approach detects a slight dip but the evidence is not strong. With the multiplatform approach, we see strong evidence of a deletion in that area. The gray shaded area indicates the CNV region identified by the HapMap 3 project release 3 (downloaded from

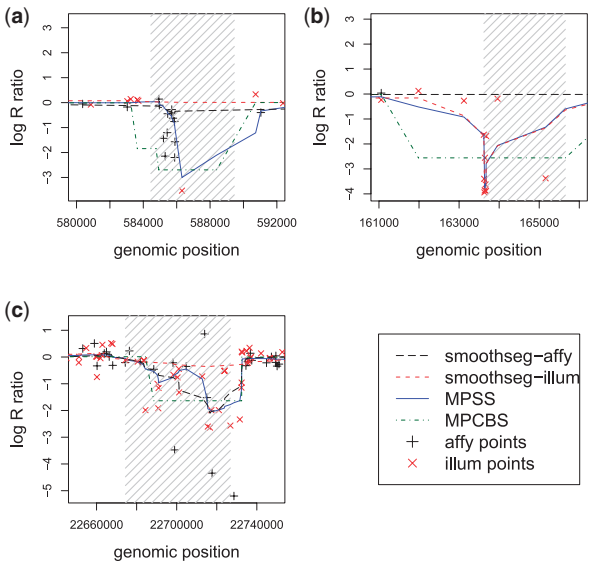


Fig. 3. Examples of segments detected by the multiplatform methods. (**a**) A deletion in Chromosome 8 of individual NA19139. Single platform smoothseg on Illumina platform was unable to identify the deletion due to lack of probes in the region. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to insufficient signal. (**b**) A deletion in Chromosome 16 of individual NA19139. Single platform smoothseg on Affymetrix platform was unable to identify the deletion due to complete lack of probes in the region. (**c**) A deletion in Chromosome 22 of individual NA19139.

ftp://ftp.ncbi.nlm.nih.gov/hapmap/cnv_data/hm3_cnv_submission.txt on 20 July 2010); this particular individual NA19139 was found to have a homozygous deletion in this region. In some cases, a single platform is unable to detect the CNV due to complete lack of probes in that region (Fig. 3b).

When signals from different platforms are inconsistent, it is difficult for the multiplatform method to detect the CNV. Even if the CNV segments are identified, they are likely to be false positives. For example, at the FDR threshold of 1e-6, the true discovery rate for the non-discrepant segments is 15.5% but it is 5.1% for the discrepant segments. On average, we remove 22 discrepant segments per individual.

Figure 4 plots the proportion of bases that overlap with Conrad's CNVs as a function of total length of CNVs for MPSS and MPCBS. MPSS has a higher proportion of base overlap with Conrad's CNVs as compared to MPCBS. Figure 5, which plots the amount of overlapping bases as a proportion of Conrad's CNVs versus the amount of overlapping bases as a proportion of each method's CNVs, also shows the better performance of MPSS as compared to all the other methods.

3.6 Application: breast cancer data

To demonstrate the applicability of the method for large studies, we apply the method to samples from the Cancer Hormone Replacement Epidemiology in Sweden (CAHRES) study, a population-based study which includes women aged 50–74 years, born in Sweden and resident there between October 1, 1993 and March 31, 1995 (Li *et al.*, 2008). A subset of 804 subjects were selected for genotyping on the HumanHap300 and HumanHap240S arrays. Clinical information

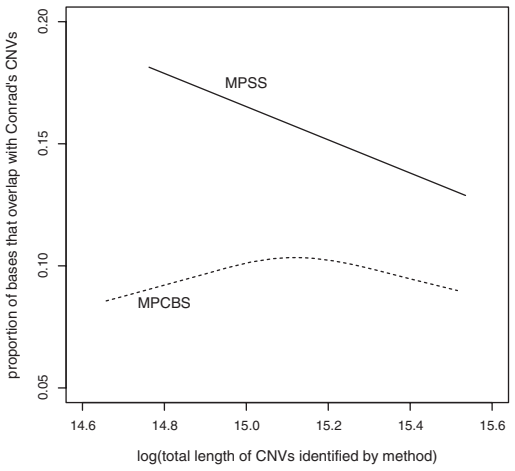


Fig. 4. Proportion of bases that overlap with Conrad's CNVs as a function of of the total length of CNVs identified by the method. A higher proportion of overlap indicates better performance.

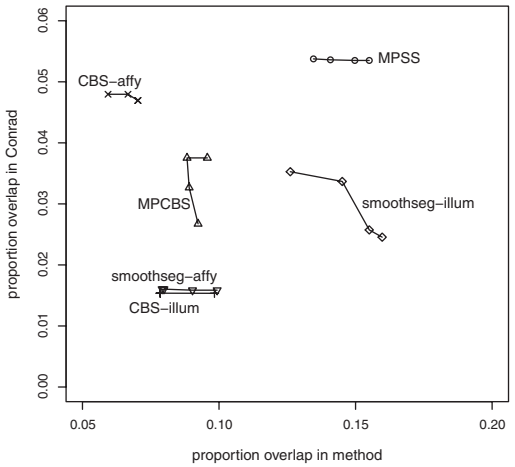


Fig. 5. The number of overlapping bases as a proportion of Conrad's CNVs and as a proportion of each method's CNVs; the different points for each method correspond to the different thresholds. A higher proportion of overlap indicates better performance.

made available to us includes lymph node status, tumor size and histologic grade. Prior to combining data from the two platforms, we use rsmooth package(http://www.meb.ki.se/∼yudpaw) to perform background correction with the smoothing parameter set to $\lambda = 10^5$.

The background-corrected intensity data is then used as inputs to MPSS algorithm. We choose the optimal smoothing parameter, λ based on the AIC criterion. For convenience, the segmentation threshold is fixed at the 5th and 95th percentile of the intensities for deletions and duplications, respectively. These are similar to objectively chosen values in the previous examples. We further filter out segments with FDR more than 0.01, number of probes less than 10, length of segments less than 1 kb and segments with discrepant signals from the two platforms.

An average of 14 deletions and 4.5 duplications are identified per individual. The median length of deletions is 113 kb and that for duplications is 140 kb. We use the method in Teo *et al.* (2010) to form common CNV segments, defined as segments with consecutive probes where there is at least 0.5% of the subjects (∼4 subjects) whose individual segments overlap with the probes. We identified 942 common segments (median length of 114.5 kb).

We test each segment for association with tumor size ($n = 540$ with size <2 cm versus $n = 60$ with size >3 cm), lymph node status ($n = 242$ lymph-node positive versus $n = 561$ negative) and tumor grade ($n = 118$ grade-1, $n = 377$ grade-2 and $n = 308$ grade-3). Fisher's exact test is used to compute the P-values. There are no significant associations with tumor size. For lymph-node status, 6 segments have $P < 0.01$ (see Supplementary Table T2). Of notable interest is segment 159 in Chromosome 3 which overlaps with the protein tyrosine phosphatase, receptor type, G (PTPRG) gene; overexpression of PTPRG was found to inhibit anchorage-independent growth and proliferation of breast cancer cells (Shu *et al.*, 2010). Another interesting segment is segment 845 in Chromosome 17, which overlaps with the ITGB4 gene, where studies have shown its expression to be correlated with tumor size and nuclear grade (Diaz *et al.*, 2005) and significantly association with basal-like breast cancer (Lu *et al.*, 2008).

Ten segments are associated with tumor grade (see Supplementary Table T3). Segment 548 in Chromosome 9 overlaps with TPM2 gene, where its protein products were found to be differentially expressed between tumor and non-tumor forming breast cancer cell lines (Harris *et al.*, 2002). Segment 691 in Chromosome 11 overlaps with the PKNOX2 gene, previously shown to be deleted in breast cancer (Issei *et al.*, 2001).

The 240K array was designed to supplement the 300K array, hence the probes on the two arrays are non-overlapping. The validation of the method in the earlier sections was performed on Affymetrix 6.0 and Illumina 1M arrays which have overlapping probes. Here, we are interested to know if the algorithm works for non-overlapping platforms. However, we do not have a 'gold standard' for CNVs of these individuals to make comparisons with. Instead, we take a random sample of non-overlapping 240 000 and 300 000 probes from the Illumina 1M platform for the 9 HapMap samples and make comparisons with the reference CNVs in the same way as before. The true discovery rate and sensitivity for the multiplatform approach is higher than that of the single platform approach: true discovery rate of 0.29 for the multiplatform approach, 0.24 for the 300K array and 0.22 for the 240K array. Sensitivity of 0.027 for the multiplatform approach, 0.027 for the 300K array and 0.017 for the 240K array.

4 DISCUSSION

We have described a new method for identifying CNVs by using data from multiple platforms simultaneously. This method allows researchers to come to a formal consensus result when data from different platforms but for the same individuals are available. The model assumes a random-effects parameter that is common to all platforms, meaning that each platform is assumed to have the same underlying copy-number pattern. We also develop an objective method to segment the estimated random effects parameter (which describes the underlying copy-number pattern) into discrete segments. In addition, we provide a method for calculating a P-value associated with a segment of interest. The P-value would indicate how likely that the segment is a deletion/duplication, and is useful for filtering out likely false positives.

Background correction is needed to make the data from the different platforms comparable; we use a robust smoother that assumes the majority of each chromosome has normal copy-number. While this assumption is likely to be true for germ-line samples, it may not hold for cancer/tumor samples. Recently, Bengtsson *et al*. (2009) developed a normalization method to bring data from different platforms to the same scale. The method uses a technique based on principal curves to estimate the normalization functions. This method was tested on data from The Cancer Genome Atlas Research Network and seems to work well on tumor samples where there is sufficient deletions and duplications in the genome, but we found that it did not work well with the germ-line samples we use. When we performed Bengtsson *et al.*(2009)'s normalization on our samples, the correlation in the copy-number estimates between the platforms increased only very slightly after normalization (see Supplementary Fig. S1 and Table T3). This could be due to insufficient CNVs in the data for the principal curves to be identified.

We illustrate the performance of MPSS using real and simulated data sets. In the comparisons using real datasets, we show that MPSS CNVs has greater amount of overlap with the reference as compared to the other methods. In the comparisons using simulated datasets, we show that the proposed method can achieve high sensitivity and specificity at reasonable noise levels.

In general, for all methods, the proportion of overlapping bases with the highly comprehensive CNV map published by Conrad *et al.* (2010) is low. However, we believe it is due to the limitation of the SNP arrays rather than the inadequacy of the algorithms. This was also noted by Zhang *et al.*(2010) where the authors investigated and found that in the regions where the reference CNVs lie, both Affymetrix and Illumina platforms do not have a shift in the intensities and hence the algorithm would not pick out the region as a CNV. Moreover, we do not know if the reference list we have used can be considered the gold standard, since it is not likely to have 100% sensitivity and specificity. Even sequencing methods only show between 72.2% and 96.5% specificity (Xie *et al*., 2009). The arrival of higher density arrays, for example, the Illumina HumanOmni2.5 and HumanOmni5 arrays will likely improve the sensitivity of CNV identification.

Another kind of multiplatform problem arises when there is some stratification of cohorts by chips; for example, if the cases and controls were typed on different chips. Differential sensitivity or false positive rates between the platforms will lead to confounding bias in the case–control comparisons. The method presented here assumes that the data from the different platforms are available for each individual, hence the algorithm could not address this problem. This is an important and valid concern and warrants further investigations.

ACKNOWLEDGEMENT

The authors thank Drs Per Hall, Kamila Czene, Liu Jianjun and Li Jingmei for providing the Swedish breast cancer data.

Funding: Swedish Science Council and National University of Singapore Start-up (Grant No. R-186-000-103-133). National University of Singapore Graduate School for Integrative Sciences and Engineering (NGS) Scholarship (to S.M.T.).

Conflict of Interest: none declared.

REFERENCES

Bengtsson,H. *et al*. (2009) A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics*, **25**, 861–867.

Benjamini,Y. *et al*. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Brent,R.P *et al*. (1973) *Algorithms for Minimization Without Derivatives.* Prentice-Hall, Englewood Cliffs, NJ.

Conrad,D.F. *et al*. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**.

Curtis,C. *et al*. (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics*, **10**, 588.

Diaz,L.K. *et al*. (2005) Beta4 integrin subunit gene expression correlates with tumor size and nuclear grade in early breast cancer. *Mod. Pathol*, **18**, 1165–1175.

Dongarra,J.J. *et al*. (1979) *LINPACK Users' Guide*. SIAM, Philadelphia.

Harris,R.A. *et al*. (2002) Cluster analysis of an extensive human breast cancer cell line protein expression map database. *Proteomics*, **2**, 212–223.

Huang,J. *et al*. (2007) Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, **23**, 2463–2469.

Huang,J. *et al*. (2009) Classification of array CGH data using smoothed logistic regression model. *Stat. Med.*, **28**, 949–951.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Issei,I. *et al*. (2001) Identification and characterization of human PKNOX2, a novel homeobox-containing ene. *Biochem. Biophys. Res. Commun.*, **287**, 270–276.

Li,J. *et al*. (2008) A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Res. Treat.*, **126**, 717–727.

Lu,S.*et al*. (2008) Analysis of integrin beta4 expression in human breast cancer: association with basal-like tumors and prognostic significance. *Clin. Cancer Res.*, **14**, 1050–1058.

McCarrol,S.A *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, **40**, 1166–1174.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pawitan,Y. (1996) Automatic estimation of coherence of bivariate time series. *Biometrika*, **83**, 419–432.

Pawitan,Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Shu,S.T. *et al*. (2010) Function and regulatory mechanisms of the candidate tumor suppressor receptor protein tyrosine phosphatase gamma (PTPRG) in breast cancer Cells. *Anticancer Res.*, **30**, 1937–1946.

Teo,S.M. *et al*. (2010) Identification of recurrent regions of copy-number variants across multiple individuals *BMC Bioinformatics*, **11**, 147.

The Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Wang,J. *et al*. (2009) The diploid genome sequence of an Asian individual. *Nature*, **456**.

Xie,C. and Tammi, M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**:80.

Zhang,N.R. *et al*. (2010) Joint estimation of DNA copy number from multiple platforms. *Bioinformatics*, **26**, 153–160.

Supplementary information for

Multi-platform Segmentation for Joint Detection of Copy Number Variants

Shu Mei Teo, Yudi Pawitan, Vikrant Kumar, Anbupalam Thalamuthu, Mark Seielstad, Kee Seng Chia and Agus Salim

Table T1: HapMap sample ID and population

ID	Population
NA10851	CEU
NA12044	CEU
NA12056	CEU
NA12057	CEU
NA18632	CHB+JPT
NA18971	CHB+JPT
NA19137	YRI
NA19138	YRI
NA19139	YRI

Table T2: Regions associated with lymph-node status

RegID	Chr	Start	End	Length(kb)	CNV type
134	2	241588064	241694279	106.2	del
159	3	61836554	61935672	99.1	del
160	3	71621386	71691966	70.6	del
348	6	14684833	14729332	44.5	del
393	6	137998026	138029601	31.6	del
845	17	71257965	71418529	160.6	del

Table T3: Regions associated with tumor status

RegID	Chr	Start	End	Length(kb)	CNV type
96	2	64511726	64541950	30.2	del
368	6	65179566	65463971	284.4	del+dup
535	9	11609839	12185026	575.2	del+dup
548	9	35504653	35813009	308.4	del+dup
615	10	92237989	92460757	222.8	del
626	10	132460162	132505487	45.3	del
691	11	124523950	124615799	91.9	del
713	12	26211854	26258097	46.2	del
770	15	86475519	86569433	93.9	del+dup
812	16	88813	152362	63.6	del+dup

Figure S1: Log R Ratio from Illumina 1M versus Log R Ratio from Affymetrix 6.0 before and after normalization for individual NA19139.

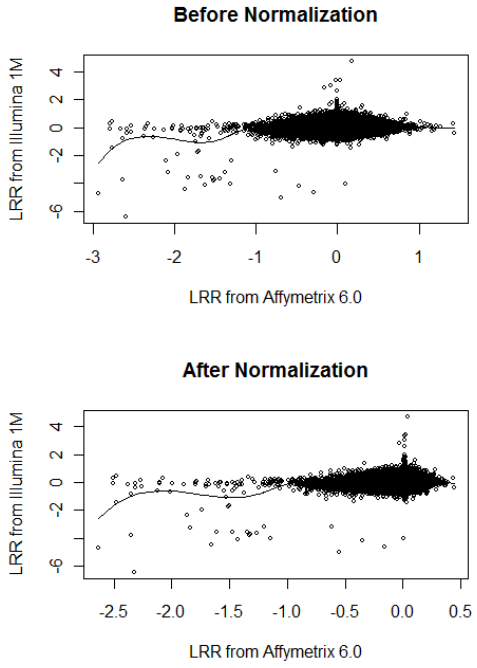


Table T3: Correlation* of the log R ratios from the Affymetrix 6.0 and Illumina 1M before and after normalization**.

Subject ID	Before normalization	After normalization
NA19139	0.051	0.072
NA15510	0.027	0.030
NA18632	0.020	0.055
NA19137	0.028	0.044
NA12057	0.021	0.022
NA18971	0.033	0.042
NA10851	0.045	0.047
NA19138	0.021	0.024
NA12044	0.016	-0.018
NA12056	-0.008	-0.009

* The correlation is calculated based on the probes that are common to both Illumina1M and Affymetrix 6.0.
** Normalization done using Bengtsson, H. *et al*.(2009)

Journal of Human Genetics (2012) 57, 400

www.nature.com/jhg

CORRIGENDUM

Regions of homozygosity in three Southeast Asian populations

Shu-Mei Teo, Chee-Seng Ku, Agus Salim, Nasheen Naidoo, Kee-Seng Chia and Yudi Pawitan

Journal of Human Genetics (2012) **57,** 400; doi:10.1038/jhg.2012.51

Correction to: *Journal of Human Genetics* (2012) **57**, 101–108; doi:10.1038/jhg.2011.132; published online 1 December 2011

The authors of the above article noted an error in publication of this paper in Figure 1. The correct figure is shown below.

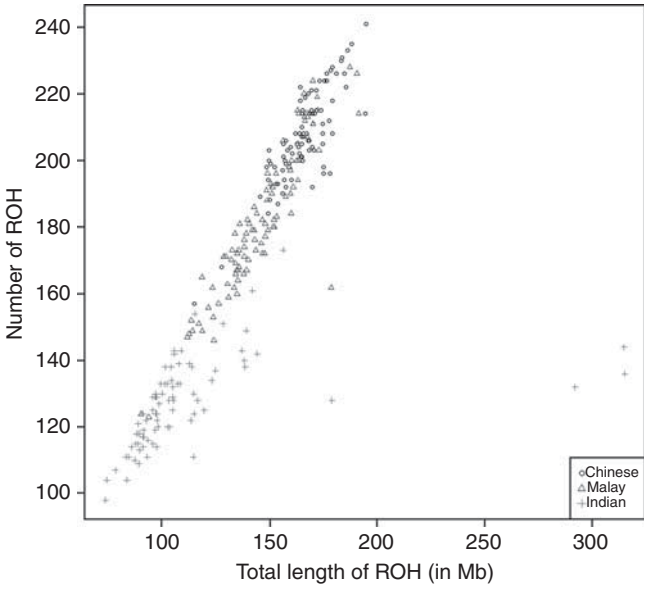


Figure 1 Number of ROH versus total length of ROHs in each individual.

The authors would like to apologize for the error. This correction does not affect the rest of the results and their interpretation as discussed in the paper.

Journal of Human Genetics (2012) 57, 101–108

www.nature.com/jhg

ORIGINAL ARTICLE

Regions of homozygosity in three Southeast Asian populations

Shu-Mei Teo[1,2,3,4,5], Chee-Seng Ku[1,2,5], Agus Salim[2], Nasheen Naidoo[1], Kee-Seng Chia[1,2,4] and Yudi Pawitan[4]

The genomes of outbred populations were first shown in 2006 to contain regions of homozygosity (ROHs) of several megabases. Further studies have also investigated the characteristics of ROHs in healthy individuals in various populations but there are no studies on Singapore populations to date. This study aims to identify and investigate the characteristics of ROHs in three Singapore populations. A total of 268 samples (96 Chinese, 89 Malays and 83 Indians) are genotyped on Illumina Human 1 M Beadchip and Affymetrix Genome-Wide Human SNP Array 6.0. We use the PennCNV algorithm to detect ROHs. We report an abundance of ROHs ($\geqslant$500 kb), with an average of more than one hundred regions per individual. On average, the Indian population has the lowest number of ROHs and smallest total length of ROHs per individual compared with the Chinese and Malay populations. We further investigate the relationship between the occurrence of ROHs and haplotype frequency, regional linkage disequilibrium (LD) and positive selection. Based on the results of this data set, we find that the frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis, demonstrating that ROHs contain information on population structure and the demographic history of a population.
Journal of Human Genetics (2012) **57,** 101–108; doi:10.1038/jhg.2011.132; published online 1 December 2011

Keywords: PennCNV; regions of homozygosity; Singapore; Southeast Asian populations

INTRODUCTION

A region of homozygosity (ROH) is defined as a continuous stretch of DNA sequence without heterozygosity in the diploid state. All genetic variations such as single-nucleotide polymorphisms (SNPs) or microsatellites within the homologous DNA segments have two identical alleles that create homozygosity.[1] Currently, there is no consensus or standardized criteria to define an ROH. Previous studies focused on ROHs larger than 1 Mb which could have led to an underestimation of the true extent of homozygosity in the human genome,[2,3] whereas more recent studies define an ROH at a minimum length of 500 kb,[4] with the intention of avoiding this underestimation. This is of relevance as shorter ROHs are now also thought to be associated with complex phenotypes.[4]

The genomes of outbred populations were first shown in 2006 to contain ROHs of several megabases.[2,3,5] Their location is markedly nonrandom, where different individuals share similar region boundaries. Some loci are caused by a single common haplotype, whereas others are a consequence of several common haplotypes that could be markedly disparate.[6] Several mechanisms for the occurrence of ROHs have been suggested, including uniparental isodisomy (a chromosomal abnormality where a child inherits two identical copies of a chromosome from one parent and none from the other) and autozygosity (where a child inherits the same common ancestral haplotype chromosomal segment from both parents). Studies have found no significant violation of Mendelian transmission in these areas and concluded autozygosity as the most likely cause for the majority of ROHs observed.[7,8]

Previous studies have also investigated the population characteristics of ROHs in healthy individuals[9–11] and performed association analyses to identify ROHs that are associated with complex diseases and traits using a case–control study design.[4,12,13] However, the majority of these studies are conducted on European populations, and only a few on Asian populations. This study aims to identify and characterize ROHs in three Singapore populations, and to investigate their relationship to linkage disequilibrium (LD), haplotype frequency and positive selection.

[1]Centre for Molecular Epidemiology, National University of Singapore, Singapore; [2]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore and [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[5]These authors contributed equally to this work.
Correspondence: S-M Teo, NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore.
E-mail: g0801862@nus.edu.sg
or Dr Y Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden.
E-mail: Yudi.Pawitan@ki.se
Received 21 April 2011; revised 30 September 2011; accepted 24 October 2011; published online 1 December 2011

MATERIALS AND METHODS

Data

We use data from the Singapore Genome Variation Project (SGVP),[14] where a total of 292 DNA samples (consisting of 99 Chinese, 98 Malays and 95 Indians) are genotyped using the Illumina Human 1 M Beadchip and the Affymetrix Genome-Wide Human SNP Array 6.0. The characteristics of copy number variations of these populations have been investigated and reported.[15] The Chinese, Indians and Malays in Singapore descended from immigrants from neighboring countries such as China (mainly from southern provinces such as Fujian and Guangdong), India (majority from south-eastern India), Indonesia and Malaysia. The detailed information on the sources of DNA samples, demographic data of the samples, sample selection, and the origin and migration history of the three Singapore populations have been described in previous publications.[14,15] A total of 268 samples (consisting of 96 Chinese, 89 Malays and 83 Indians) are used in the subsequent analysis after removing samples on the basis of high rates of SNP missingness (greater than 2%), excessive heterozygosity or cryptic relatedness by excessive identity-by-states. Population membership is ascertained on the basis that all four grandparents belong to the same population, and samples that display either evidence of admixture or clear evidence of discordance between self-reported and genetically inferred population membership are excluded.

SNP genotypes are obtained from the SGVP website (http://www.nus-cme.org.sg/sgvp/). These SNPs have undergone a series of quality control measures,[14] including removing SNPs with SNP missingness $\geqslant 5\%$ and *P*-value <0.001 for a test of departure of Hardy–Weinberg Equilibrium (HWE), resulting in ~ 1.58 million SNPs per population remaining. Quality control measures were conducted seperately for each of the populations.

Identification of individual-specific ROHs

Individual-specific ROHs are identified using the PennCNV algorithm[16] for the Illumina and Affymetrix arrays based on the log R ratio and B allele frequency for each sample. The ROHs identified by PennCNV are copy neutral events, meaning that one-copy deletions are excluded. We exclude regions <500 kb. To further filter regions that may be called erroneously by PennCNV, we check the SNPs genotypes for the number of heterozygous genotypes within the region. Ideally, we would expect no heterozygous genotypes in the region, but we allow for some heterozygosity that may be due to genotyping errors or other causes.

We investigate the effect of allowing some heterozygosity on the relationship between ROH and LD. From a simulation (see Supplementary Methods, 'Simulation' section), we observe that ROH detection is very sensitive to heterozygosity present either due to mutation or genotyping errors, whereas the LD in the region is largely preserved despite the mutations introduced. By not allowing any heterozygosity, we miss detecting older ROHs in many individuals and this affects the formation of the common regions. So, to capture the LD/haplotype structure using ROHs, it is important to allow a small percentage of heterozygosity.

We use a binomial probability upper bound to calculate a confidence score for each region (see Supplementary Methods, 'Confidence scores calculation' section). The confidence score takes into account the amount of heterozygosity, as well as the SNP density, and is an indication of how confident we are that the ROH is true. In general, the confidence scores for regions detected by the Affymetrix platform are lower than that detected by the Illumina platform (see Supplementary Methods Figure S1). We decide to use the Illumina platform with more than 1 million SNPs for ROHs detection but still use the combined genotypes from 1.58 million SNPs from both platforms in the calculation of confidence scores. Several summary statistics are computed to describe and compare the characteristics of ROHs in the three Singapore populations.

Identification of common ROHs

We identify common ROH loci using a previously published method.[17] We define common loci as regions with consecutive probes where at least 5% of the subjects (that is, 13 subjects) have individual regions that overlap with the probes. Occasionally, individual regions within a common locus can show considerable variations in their boundaries, resulting in a heterogeneous region. To refine the identified common loci, we form clusters of regions by requiring all individual regions within a cluster to overlap by at least 80%. For each common locus, individual regions are said to be concordant if it overlaps with at least 80% of the length of the locus. Common loci with <2 concordant individuals or <500 kb or having a SNP density <0.2 (SNP per kb) are discarded. The common loci are further refined as the intersection of the concordant regions. We perform population comparisons and test of departure of HWE for each locus. For each set of tests, we account for multiple comparisons using the false discovery rate,[18] with results or discoveries considered interesting at false discovery rate of <0.01.

Quantification of regional LD

The two most widely used measures to quantify the amount of LD between two markers are the D' and r^2 statistics.[19] Here, instead of LD between two markers, we are interested in the amount of LD in a region. For all SNPs in a region, we calculate the pairwise D' (and r^2). We perform eigenvalue decomposition on the D' (r^2) matrix and calculate the percentage explained by the first eigenvalue (y). This percentage will take values between $100/n$ and 100, where n is the number of (polymorphic) SNPs in the region. To make the percentages comparable across regions with different number of SNPs, we scale it such that the value varies between 0 and 1. So, $y^*=(y-100/n)/(100-100/n)$. The higher the value of y^*, the stronger the LD in the region.

Haplotypes in ROH loci

For each common locus, we use phased genotypes (using the program fastPHASE version 1.3, see Supplementary Methods in Teo *et al*.[14] for details on the choice of parameters for phasing) to determine the different haplotypes present in the three populations. To reduce the dimension of the data, we consider only the top three most frequent haplotypes and combine the others as 'other haplotypes', that is, we categorize each region into four alleles (top three most common haplotypes and 'other' haplotypes). Each individual has two alleles for each region. For convenience, we will refer to the alleles as A, B, C and D.

Identification of regions with differential LD between populations

We use a previously published method, varLD,[20,21] to identify regions with differential LD between populations. Briefly, the method tests for equality between two LD matrices for a user-defined window size, shifting each window one SNP at a time. We calculate the varLD score for a window size of 50 SNPs for the signed r^2 matrices.[21] For each pair of populations, a region is considered to have differential LD if consecutive positions are above the 95th percentile of the genome-wide varLD score. We restrict to regions >500 kb for comparison with ROHs. We exclude the region if it overlaps by $>50\%$ with copy number variations previously reported for the same set of individuals,[14] as LD measures for regions that encapsulate copy number variations may not be reliable.[21]

RESULTS

Summary statistics of individual ROHs

We discard regions whose confidence scores are below the 25th percentile of the confidence scores. Table 1 summarizes the characteristics of ROHs. On average, the Indian population has lower number of ROHs compared with the Chinese and Malay populations. There is wide inter-individual difference in the number of ROHs, which ranges from 98 (sample 334_01 and 461_01) to 241 (sample 81_01). More than 80% of the ROHs are <1 Mb in length. The largest ROH spans a length of ~ 68.5 Mb, and is detected in one Indian individual (sample 408_01) in Chromosome 3. A total of 32 ROHs larger than 10 Mb are detected (Table 2). Interestingly, three Indian samples (397_01, 290_01 and 408_01) have five or more of these 'extremely long' ROHs. Figure 1 plots the number of ROHs versus the total length of ROHs in each individual. We see clusters of the three populations, indicating that number and length of ROHs differ among populations. This result was also observed by Kirin *et al*.[22]

Table 1 Characteristics of ROHs in three Singapore populations (using 1 029 591 SNPs from the Illumina 1 M platform)

Characteristics	*Chinese (n=96)*	*Malay (n=89)*	*Indian (n=83)*
Number of ROHs per individual			
Mean	207	179	126
Median	206	178	127
Minimum	157	123	98
Maximum	241	228	173
Length of ROHs (in kb)			
Mean	800.9	806.1	879.6
Median	670.5	672.8	666.3
Maximum	23 230	21 850	68 500
Total length of ROHs per individual (in Mb)			
Mean	166.1	144.6	111.2
Median	165.7	143.4	100.5
Minimum	115.4	90.49	73.91
Maximum	195.4	191.9	315.6
Size distribution of ROHs (proportion, %)			
500kb–1 Mb	83.0	82.5	83.5
$\geqslant$1 Mb	17.0	17.5	16.5

Abbreviations: ROHs, regions of homozygosity; SNPs, single-nucleotide polymorphisms.

Summary statistics of common ROHs

We identify 1256 common ROH loci in all three populations (Supplementary Table 1), where 90% of the loci overlap with UCSC genes (http://genome.ucsc.edu/), and 292 (23%) overlap with genes listed in the Online Mendalian Inheritance in Man Morbid Map (ftp://ftp.ncbi.nih.gov/repository/OMIM/ARCHIVE/morbidmap). For each locus, we test for differences among the three populations in terms of ROH frequencies and haplotype frequencies, and 47 loci ($<$4%) differ significantly in frequencies while 899 loci (69%) differ significantly in haplotype frequencies among the populations. Approximately 52% of the loci are detected in $>$5% (more common ROH loci) of individuals (Figure 2). Figure 3 shows the length distribution of the ROH loci; $\sim$78% of the ROH loci are $\leqslant$1 Mb, and majority of the long ROH loci ($>$1Mb) are in the range of 1–2 Mb. The proportion of the genome that is in the different ROH length categories differs among the three populations (Figure 4). The Chinese and Malays have more ROHs of shorter lengths compared with the Indians, while the Indians have more ROHs in the longer length categories ($>$4 Mb).

We compare the common loci we found to that published in previous studies.[10,23] Two regions are defined to overlap if the regions have a reciprocal overlap of at least 50%. Nothnagel *et al*.'s study[10] surveys ROHs in Europeans; we found that all 10 regions listed as 'ROH islands' (meaning they have a high population frequency) in their study overlap with an ROH loci found in this study, suggesting

Table 2 ROHs larger than 10 Mb

Chromosome	*Start*	*End*	*Length*	*Sample*	*Ethnicity*
1	120 837 663	143 420 875	22 583 213	108_01	Chinese
6	3 217 193	26 449 280	23 232 088	17_01	Chinese
16	34 034 376	45 968 704	11 934 329	131_01	Chinese
1	120 837 663	143 420 875	22 583 213	218_01	Chinese
8	41 842 707	52 102 021	1 025 9315	465_01	Malay
1	94 915 135	108 531 282	13 616 148	174_01	Malay
11	19 924 676	41 772 573	21 847 898	174_01	Malay
1	67 073 684	90 862 713	23 789 030	290_01	Indian
3	175 758 479	190 839 635	15 081 157	290_01	Indian
4	41 334 756	55 223 410	13 888 655	290_01	Indian
6	112 768 454	147 227 544	34 459 091	290_01	Indian
11	86 911 515	131 748 067	44 836 553	290_01	Indian
14	71 970 357	88 634 741	16 664 385	290_01	Indian
17	152 362	10 559 477	10 407 116	290_01	Indian
3	102 253 981	170 758 820	68 504 840	408_01	Indian
6	79 661	14 549 208	14 469 548	408_01	Indian
6	121 892 269	132 743 942	10 851 674	408_01	Indian
13	52 267 631	77 893 750	25 626 120	408_01	Indian
13	78 497 109	94 452 168	15 955 060	408_01	Indian
22	37 722 142	49 582 267	11 860 126	408_01	Indian
13	74 778 668	100 318 094	25 539 427	367_01	Indian
3	158 294 635	169 108 914	10 814 280	361_01	Indian
6	90 065 419	106 409 967	16 344 549	397_01	Indian
7	28 668 234	43 132 968	14 464 735	397_01	Indian
7	80 118 053	105 839 742	25 721 690	397_01	Indian
8	590 729	10 908 015	10 317 287	397_01	Indian
9	33 415 385	45 059 163	11 643 779	397_01	Indian
9	66 448 030	100 731 809	34 283 780	397_01	Indian
10	121 636	24 722 946	24 601 311	397_01	Indian
15	63 308 076	73 720 143	10 412 068	397_01	Indian
7	9 983 924	21 830 396	11 846 473	76_01	Indian
13	47 337 000	72 862 520	25 525 521	76_01	Indian

Abbreviations: ROHs, regions of homozygosity.

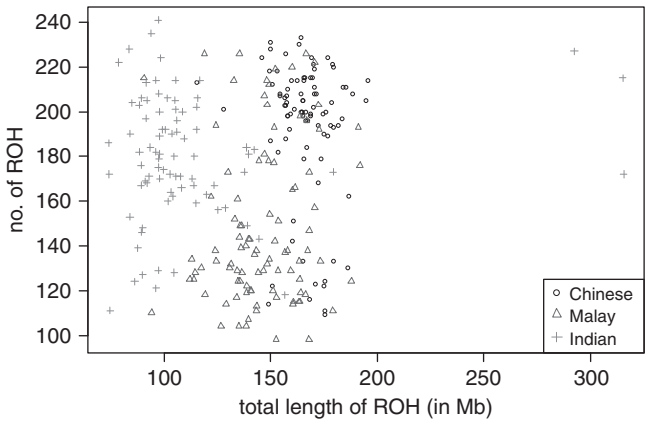


Figure 1 Number of ROH versus total length of ROHs in each individual. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

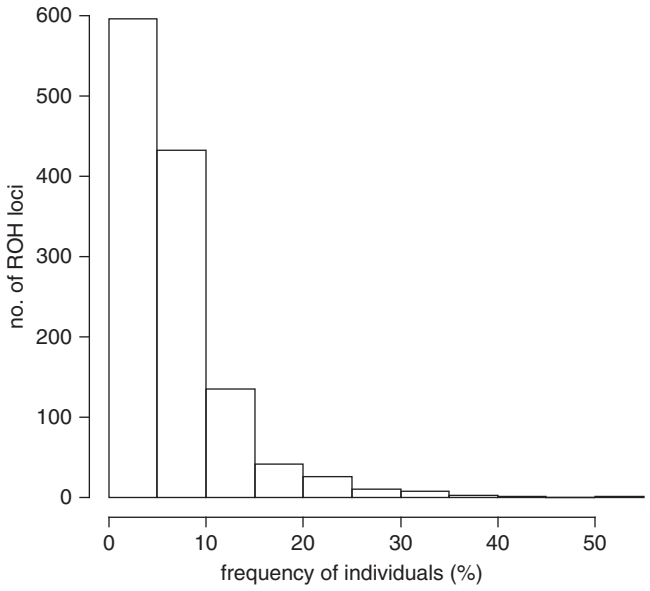


Figure 2 Number of ROH loci in the respective population frequency classes.

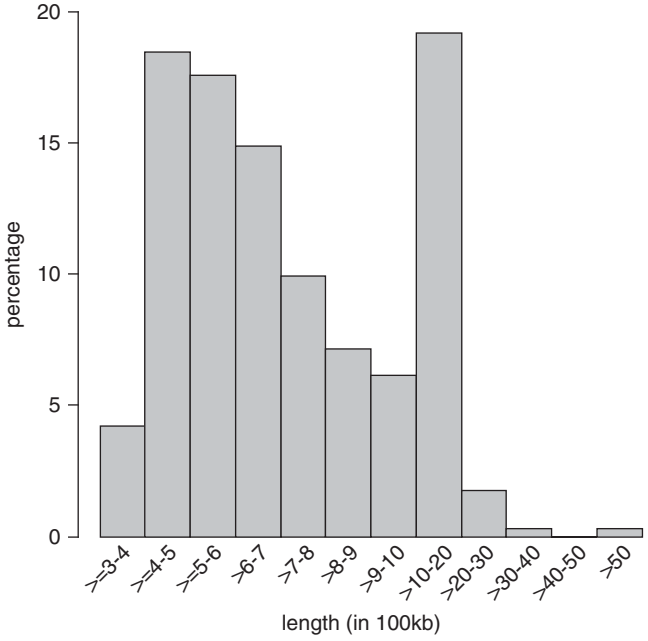


Figure 3 Percentage of ROH loci in the respective length classes.

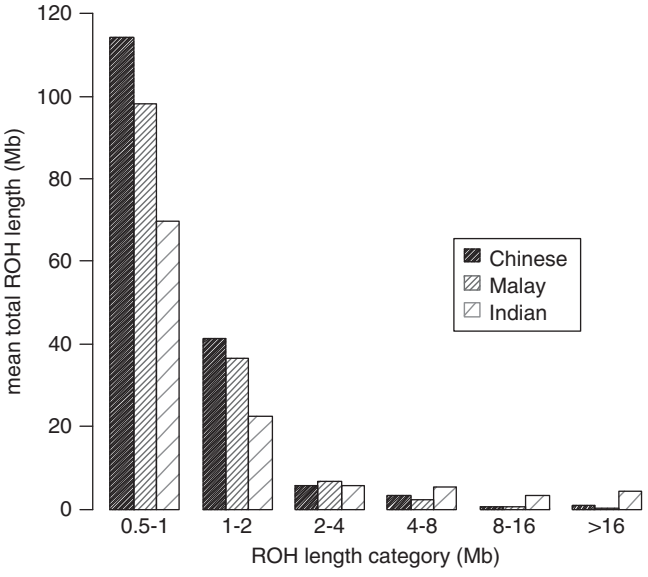


Figure 4 Percentage of ROH loci in the respective length classes. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

that these regions are not specific to Europeans (see Supplementary Methods Table S1). The population frequencies of these ROHs in our populations differ from that reported in Nothnagel *et al*.'s study,[10] but formal testing is inappropriate as the methods used to calculate the frequencies are different.

Auton *et al*.'s study[23] surveys ROHs in Mexicans, Europeans, East Asians and South Asians; we found that out of 34 high-frequency ROHs (defined as being found in at least 10% of individuals within a population) 11 overlap with an ROH locus found in our study (see Supplementary Methods Table S2). All the regions that overlap are found in the East Asian population, except for one region in Chromosome 4, which is present in all populations. The frequencies of these ROHs are, however, quite low in our population (1–4%).

Association with haplotype frequency and regional LD

Figure 5 shows that the frequency of an ROH is positively associated with the total frequency of the top three haplotypes (correlation of 0.69), and Figure 6 shows that as the frequency of an ROH increases, so does $y*_{D'}$ and $y*_{r^2}$ (figure is shown for the Malay population, similar figures for the Chinese and Indians are shown in Supplementary Methods Figures S2 and S3). If we assume random mating, the homozygosity of any region will be high when there are few haplotypes present at high frequency, thus it reinforces autozygosity as the mechanism for the occurrence of an ROH. These empirical results suggest that there is positive correlation between the frequency of an ROH and the frequency of the common haplotypes, and also between the frequency of an ROH and LD in the region.

Frequency of ROHs and frequency of haplotypes within ROHs

To assess if there is a difference in the location and frequency of ROHs among the populations, we perform principal component analysis

(PCA) using absence/presence of the common ROH loci. For each individual, we check if that individual has an ROH that is concordant with the common ROH. We can view the matrix input for the PCA analysis as a matrix of 1's and 0's where each row corresponds to an individual and each column corresponds to a common loci, so that the (i, j) entry indicates whether individual i has a concordant ROH at locus j. From Figure 7, we see that the Indians are quite well separated from the Chinese and Malays, and that there is some separation between the Chinese and Malays. This implies that the location and frequency of occurrence of ROHs differ among populations.

However, interestingly, populations can share the same (or similar) ROH location, but the common haplotypes driving the ROH can be markedly disparate. One example is a 700-kb ROH in Chromosome 16 (location 30,438,046–31,137,964) that overlaps with the Vitamin K epoxide reductase complex subunit 1 (*VKORC1*) gene (location 31,009,956–31,013,551). Genetic polymorphisms within the gene have been found to correlate with differences in warfarin dosage and response in many studies.[24–26] In the Singapore populations, the Indians were observed to display warfarin resistance, thus requiring a higher dose as compared with the Chinese and Malays.[26–29] There is no significant difference in ROH frequencies among the populations (ROH frequencies of 21, 13 and 20% for the Chinese, Malays and Indians, respectively). However, if we examine the haplotypes in this region, there is significant difference. Fisher's exact test performed on the frequencies of the top three most frequent haplotypes results in a P-value $< 10^{-6}$. In particular, the difference in haplotype frequencies of the Indians differs markedly from the Chinese and Malays. This is highlighted in Table 3, where haplotype A dominates in the Chinese and Malays but is almost absent in the Indians, while haplotype B dominates in the Indians but is almost absent in the Chinese and Malays. Haplotypes A and B differ at 104 locations out of the 158 SNPs in this region.

We also perform PCA on the allele counts of the haplotypes as described in the section 'Haplotypes in ROH loci'. The first two components separates the Indians from the Chinese and Malays while the third component further separates the Chinese from the Malays (see Figure 8). This suggests that ROH loci contain much genetic ancestral haplotype information of a population.

Testing departure from HWE

Using the estimated frequencies of the top three haplotypes, we are able to calculate the expected frequencies of the corresponding genotypes. For the observed frequencies, we use the unphased genotypes. For each individual, we can identify the haplotypes without phase information when all the SNPs in the region are homozygous (removing SNPs where we had allowed heterozygosity in the detection

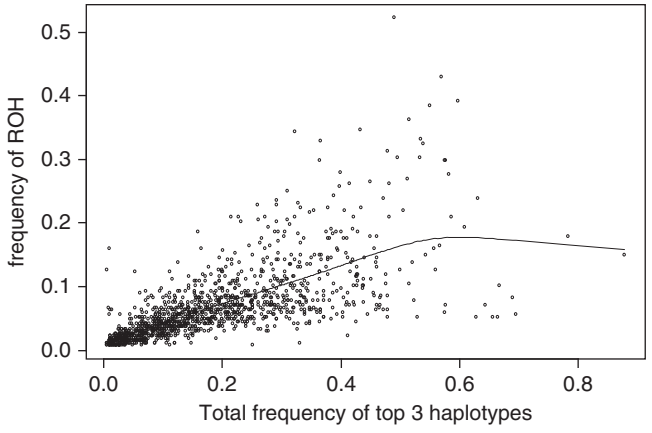


Figure 5 Frequency of ROH loci versus total frequency of top three haplotypes.

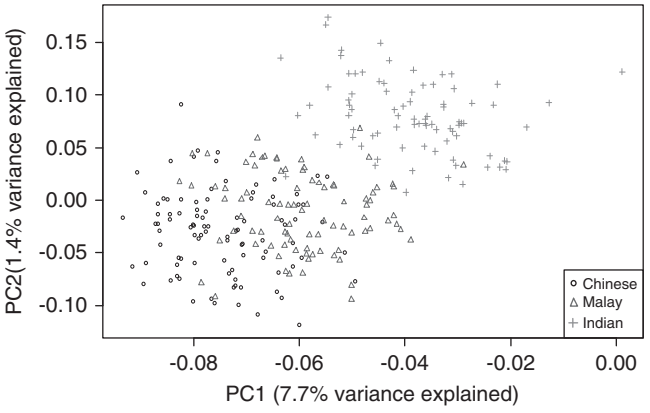


Figure 7 Principal component 2 versus principal component 1 using absence/presence of 1256 common ROHs. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

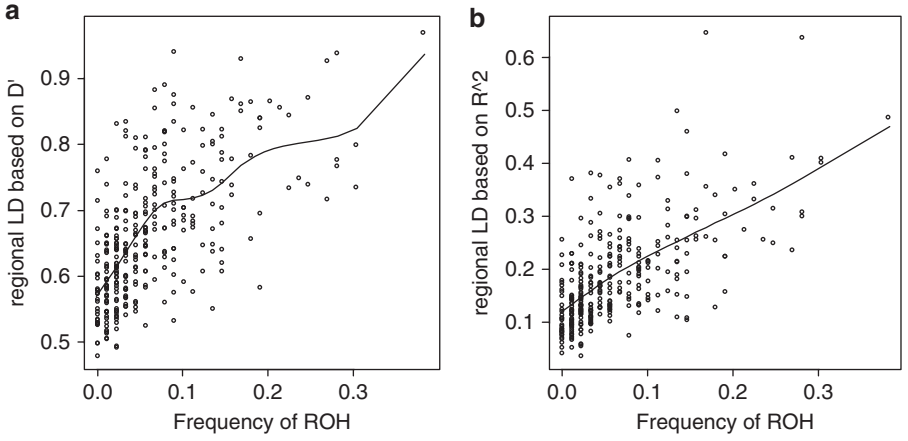


Figure 6 Regional LD versus frequency of ROH based on (**a**) D' matrix and (**b**) r^2 matrix. These results are based on the Malay population.

Table 3 Haplotype frequencies of three populations in a ROH in Chromosome 16 that overlaps with VKORC1

	Haplotype A	*Haplotype B*	*Haplotype C*
Chinese	0.31	0.0052	0.099
Malay	0.28	0.045	0.10
Indian	0.0060	0.34	0

Abbreviation: ROH, region of homozygosity.

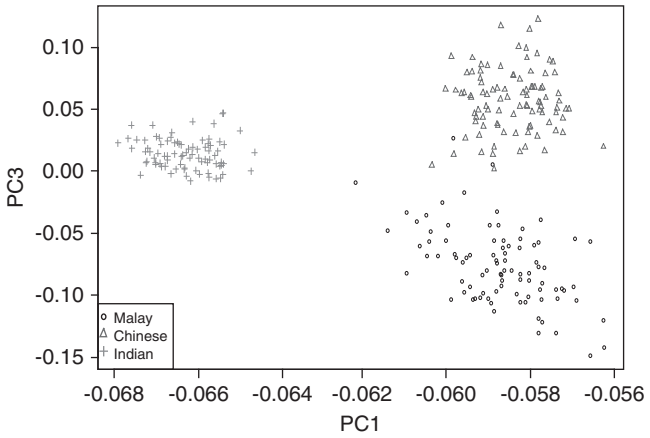


Figure 8 Results of PCA on haplotype frequencies of ROH regions. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

of ROH). With that, we are able to obtain observed frequencies for the (A, A), (B, B) and (C, C) genotypes. We use the χ^2 test with three degrees of freedom to test if there is departure of the observed from the expected. A large majority of ROH loci ($>92\%$) adhere to HWE, suggesting that assumptions of autozygosity and random mating are true for most ROH loci. Of the regions that show departure from HWE (false discovery rate <0.01), majority show excess homozygosity than would be expected. The reasons for departure from HWE are not immediately clear, and could be due to various reasons such as positive selection (see section 'Comparison with regions associated with positive selection') or nonrandom mating.

Comparison with varLD

As described in the section 'Identification of regions with differential LD between populations', we identify 16, 10 and 13 regions with differential LD variation between the Chinese and Indian populations, Malay and Indian populations, and Chinese and Malay populations, respectively. Of the 16 regions, 14 overlap with a common ROH and 10 out of 14 show significant differences in haplotype frequency between the Chinese and Indian populations. Of the 10 regions, 7 overlap with a common ROH and 7 out of 7 show significant differences in haplotype frequency between the Malay and Indian populations. Of the 13 regions, 8 overlap with a common ROH and 8 out of 8 show significant differences in haplotype frequency between the Chinese and Malay populations.

We observe that the majority of regions (74%) that show LD differences between populations correspond to regions where ROHs are observed, and furthermore, the haplotype frequencies in these regions differ between the populations. These results indicate that ROH patterns explain a large proportion of LD variations.

Comparison with regions associated with positive selection

We investigate if the regions detected for recent positive natural selection overlap with ROHs. We consider the top 10 candidate regions for recent positive selection in each of the populations, as published in a previous study.[14] These regions were detected based on the clustering of SNPs with high integrated haplotype score.[30] Out of the 30 regions considered, 28 regions overlap with a common ROH defined in this study, with 20 regions completely within an ROH and the other 8 regions with a high percentage of overlap (at least 60%). This suggests the occurrence of ROHs as a possible consequence of positive selection, where the positively selected haplotypes rise to a high frequency, resulting in a high possibility of ROHs due to autozygosity.

Out of the 28 regions, 10 of them overlap with an ROH that failed HWE. Performing Fisher's exact test on a 2 by 2 table with indicators for departure from HWE and indicators for positive selected regions as rows and columns, we obtain an odds ratio of 1.89 (P-value=0.05). The departure from HWE may be a consequence of positive selection. An ROH that has a higher frequency than would be expected for its length may also be an evidence of positive selection (see Supplementary Methods Figure S8).

Effect of heterozygosity on the relationship between ROH and LD

When we filter the individual regions using a stricter confidence threshold of the 75th percentile (that is, allowing less heterozygosity), we identify 414 common regions, but the relationship of these regions with haplotype frequency, regional LD and positive selection is weak (see Supplementary Methods Figures S4 and S5 and section Comparison with VarLD (results based on these 414 common regions)). We also see poorer separation of the populations by PCA, but this is likely due to the fewer number of common regions identified. At the 25th percentile threshold, the percentage of heterozygosity is still kept low at $<5\%$ for a large majority of the regions (See Supplementary Methods Figure S9). With an overly strict confidence score threshold, many regions are omitted and this decreases the number of common regions formed from 1256 to 414. Allowing for some heterozygosity within the regions allows detection of older ROH loci (heterozygosity caused by recent recombination or mutation), which have a stronger relationship with LD and positive selection (see Simulation section in Supplementary Methods).

DISCUSSION

In summary, this study identifies and investigates the population characteristics of ROHs in three Singapore populations, Chinese, Malay and Indian. We report an abundance of ROHs, with an average of >100 ROHs per individual. On average, the Indians have lower numbers and total length of ROHs per individual than the Chinese and Malays, possibly indicative of a larger founder population. However, there are several Indians with multiple large ROHs, suggesting that they may be offsprings of parents who are close relatives. In India, consanguineous marriages are more prevalent in the South, especially in Tamil Nadu, from where many Singapore Indians descended. From the Consanguinity/Endogamy Resource (http://www.consang.net/index.php/Main_Page), data from a 1982 study have shown the prevalence of consanguineous marriages among Singapore Indians to be 4% compared with only 0.3% in Singapore Chinese. Published data have shown that the number of ROHs of several megabases increase markedly in the offsprings of consanguineous marriages,[3,24] with an average of 6.25% homozygosity expected in the genome of the offsprings of first cousin marriages.[7] Li *et al.*[3] have shown that in a family with four children from first cousin

marriages, multiple ROHs ranging from 3.06 to 53.17 Mb were observed in all the children. Woods *et al.*[24] have also shown a marked increase in homozygosity levels in individuals with a recessive disease whose parents were first cousins, where, on average, 11% of their genomes were homozygous.

In addition, we identify 1256 common ROH loci, and investigate the occurrence of ROHs and haplotype frequency, regional LD and positive selection. Based on the results for this data set, we find that the frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The preferential occurrence of ROHs in regions of high LD and low recombination has also been observed in other studies.[10] The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with ROH loci. By considering both the location of the ROH and the allelic form of the ROH, we are able to separate the populations by PCA, demonstrating that ROHs contain information on population structure and the evolutionary and demographic history of a population.

The ability of genome-wide SNP markers for population structure analysis has been widely acknowledged. Here, we are not proposing the superiority of ROHs in population structure analysis. It is expected that using genome-wide SNP data allows very good separation of populations through PCA because of the amount of information it contains (see 14 for PCA analysis using SNPs on the same population samples). In this paper, we have shown that it is possible to distinguish populations using just $\sim$1000 segments of the genome. Comparatively, if we were to choose 1000 random segments of the genome and perform a similar analysis, we would not obtain as good a separation as with ROHs (see Supplementary methods Figure S7). The unique characteristics of ROHs allow us to study common haplotypes conveniently; it is complementary to SNP-based analysis. In SNP-based analysis, we simply compare SNP-level frequencies between populations but in ROH-based analysis, we are able to capture differences in LD or haplotype structures.

Majority of the ROH loci overlap with known genes but their association with complex phenotypes is still rudimentary. This warrants further characterization of ROHs in different populations, investigation of their roles in the genetics of complex phenotypes and further studies of population evolutionary genetics. These future studies will be of importance given the abundance of ROHs in the human genome and the differences of ROHs between populations.

A sufficiently large number of SNPs is required to accurately detect ROHs.[1,2] To this end, we have used two highly dense SNP arrays (Illumina 1 M and Affymetrix 6.0) with $>$1.58 million unique SNPs. Using a confidence score metric that takes into account percentage of heterozygosity as well as the number of SNPs in the region, we discard individual regions whose confidence scores are below the 25th percentile of the confidence scores. We use the PennCNV algorithm that relies on signal intensity data to detect putative ROHs. We then filter out false positives by checking SNP genotypes within the ROH. To our knowledge, most studies on ROHs use only SNP genotypes, but this approach may produce false positives caused by hemizygous deletions. On the other hand, due to the noise in signal intensity data, the regions called by PennCNV could also result in false-positive regions. We feel it is important to use a combination of the methods (that is, signal intensity data and genotype data) to minimize false-positive rates.

We also use PLINK, a widely used software for ROH detection, on genotypes from both platforms using the following parameters: 500 kb window with two heterozygous SNPs allowed, minimum length of 500 kb, 50 SNPs as minimum number of SNPs and minimum density of 1 SNP per 10 kb. We find that 75% of the regions found by PennCNV are detected by PLINK, suggesting that the results of the analysis will likely give similar conclusions using PLINK. A formal and systematic comparison of multiple algorithms for ROH detection will be interesting.

Potential biases in the detection of ROHs include false-negative regions due to ascertainment bias in SNP selection for the SNP arrays and false-positive regions due to the lack of minor allele frequency (MAF) criterion applied before the identification of ROHs. With regards to the former, SNPs from genotyping platforms are mostly tagged SNPs from the HapMap project, so populations that were not analyzed in the HapMap project will have less chance of their population-specific SNPs being included in the array. However, both the Illumina 1 M and Affymetrix 6.0 arrays have a high marker density and uniformity. With regards to the later, we do not expect our results to be affected considerably by not filtering SNPs with low MAF, for several reasons. First, we have very dense SNP genotyping data of $>$1.58 million SNPs, and as an ROH is defined as a region of consecutive homozygosity of $>$500 kb, it is unlikely that there exists a large number of consecutive low-MAF SNPs that cause a false-positive identification. In any case, these monomorphic/near monomorphic SNPs are uninformative and would not affect the haplotype analyses. It is of concern if the region is detected because the monomorphic/low-MAF SNPs are genotyped, whereas other SNPs present in the region are missed (due to ascertainment bias). However, as ROH detection is not reliant on a single SNP, but on many consecutive homozygous SNPs in a 500 kb region, we do not expect either issue to be of serious concern.

Some studies[23] have adopted the strategy of removing SNPs in high LD before defining an ROH (that is, thinning the data set but requiring a lower number of SNPs for the definition of ROH). However, we found poor correlation between the frequency of the ROHs we identified and the mean or median pairwise D' or r^2 statistics (for SNPs within the ROH, up to 250 kb apart, see Supplementary Methods Figure S6), meaning that a SNP being in high LD in the vicinity is not sufficient for its inclusion in an ROH, and a SNP in low LD is not sufficient for its exclusion in an ROH.

In conclusion, our study is one of the first to describe the population characteristics of ROHs in the three Singapore populations (Chinese, Malay and Indian). Our results are in support that ROHs contain population demographic and ancestral haplotype information.

ACKNOWLEDGEMENTS

We thank Dr Teo Yik Ying for helpful discussions related to this work and Rick Ong for identifying regions with differential LD between populations using the VarLD program. TSM acknowledges support from the National University of Singapore Graduate School for Integrative Sciences and Engineering (NGS) Scholarship.

1 Ku, C. S., Naidoo, N., Teo, S. M. & Pawitan, Y. Regions of homozygosity and their impact on complex diseases and traits. *Hum. Genet.* **129,** 1–15 (2011).
2 Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15,** 789–795 (2006).
3 Li, L. H., Ho, S. F., Chen, C. H., Wei, C. Y., Wong, W. C., Li, L. Y. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* **27,** 1115–1121 (2006).
4 Yang, T. L., Guo, Y., Zhang, L. S., Tian, Q., Yan, H., Papasian, C. J. *et al.* Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *J. Clin. Endocrinol. Metab.* **95,** 3777–3782 (2010).
5 Simon-Sanchez, J., Scholz, S., Fung, H. C., Matarin, M., Hernandez, D., Gibbs, J. R. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homo-

zygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16,** 1–14 (2007).
6 Curtis, D., Vine, A. E. & Knight, J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann. Hum. Genet.* **72,** 261–278 (2008).
7 Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* **65,** 1493–1500 (1999).
8 Curtis, D. Extended homozygosity is not usually due to cytogenetic abnormality. *BMC Genet.* **8,** 67 (2007).
9 McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R, Franklin, C. S., Pericic, M., Barac-Lauc, L. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83,** 359–372 (2008).
10 Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet.* **19,** 2927–2935 (2010).
11 O'Dushlaine, C. T., Morris, D., Moskvina, V., Kirov, G., Consortium, I. S., Gill, M. *et al.* Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18,** 1248–1254 (2010).
12 Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl Acad. Sci. USA* **104,** 19942–19947 (2007).
13 Nalls, M. A., Guerreiro, R. J., Simon-Sanchez, J., Bras, J. T., Traynor, B. J., Gibbs, J. R *et al*. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10,** 183–190 (2009).
14 Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E. *et al.* Singapore genome variation project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19,** 2154–2162 (2009).
15 Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31,** 851–857 (2010).
16 Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. *et al*. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).
17 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* **11,** 147 (2010).
18 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).
19 Lewontin, R. C. & Kojima, K. The evolutionary dynamics of complex polymorphisms. *Evolution* **14,** 458–472 (1960).
20 Ong, R. T. H. & Teo, Y. Y. varLD: A program for quantifying variation in linkage disequilibrium patterns between populations. *Bioinformatics* **26,** 1269–1270 (2010).
21 Teo, Y. Y., Fry, A. E., Bhattacharya, K., Small, K. S., Kwiatkowski, D. P. & Clark, T. G. Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.* **19,** 1849–1860 (2009).
22 Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., Mckeigue, P. M. & Wilson, J. F. Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5,** e13996 (2010).
23 Auton, A., Bryc, K., Boyko, A. R., Lohmueller, k.E., Novembre, J., Reynolds, A. *et al.* Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19,** 795–803 (2009).
24 Woods, C. G., Cox, J., Springell, K., Hampshire, D. J., Mohamed, M. D., McKibbin, M. *et al.* Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am. J. Hum. Genet.* **78,** 889–896 (2006).
25 Aquilante, C. L., Langaee, T. Y., Lopez, L. M., Yarandi, H. N., Tromberg, J. S., Mohuczy, D. *et al.* Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clin. Pharmacol. Ther.* **79,** 291–302 (2006).
26 Harrington, D. J., Underwood, S., Morse, C., Shearer, M. J., Tuddenham, E. G. D. & Mumford, A. D. Pharmacodynamic resistance to warfarin associated with a Val66Met substitution in vitamin K epoxide reductase complex subunit 1. *Thromb. Haemost.* **93,** 23–26 (2005).
27 Zhu, Y., Shennan, M., Reynolds, K. K., Johnson, N. A., Herrnberger, M. R., Valdes, R. Jr. *et al.* Estimation of warfarin maintenance dose based on VKORC1 (−1639 G > A) and CYP2C9 genotypes. *Clin. Chem.* **53,** 1199–1205 (2007).
28 Yuen, E., Gueorguieva, I., Wise, S., Soon, D. & Aarons, L. Ethnic differences in population pharmacokinetics and pharmacodynamics of warfarin. *J. Pharmacokinet. Pharmacodyn.* **37,** 3–24 (2009).
29 Lee, S. C. Inter-ethnic variability in warfarin requirement is explained by VKORC1 genotype in an Asian population. *Clin. Pharmacol. Ther.* **79,** 197–205 (2006).
30 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Supplementary Methods for

Regions of homozygosity in three Southeast Asian populations

Shu-Mei Teo, Chee-Seng Ku, AgusSalim, NasheenNaidoo, Kee-Seng Chia, YudiPawitan

Simulation

We perform a simulation study to investigate the effect of allowing/ not allowing heterozygosity in defining ROHs. We pick a 70 SNP ROH locus with 29 haplotypes whose observed frequencies are 31%, 23%, 16%, 10%, 6% and so on. We randomly sample a pair of haplotypes for each of 1000 individuals from the 29 haplotypes with the above mentioned frequencies. We repeat the process 100 times. Using the same definition of regional LD (based on D', see main manuscript), we obtain an average ROH frequency of 19% and an average regional LD of 81.9%.

Next, we introduce three mutations per haplotype such that the total number of haplotypes is now 58. The new haplotypes have frequencies that are 10% of their 'parent' haplotype from which they mutated from. Thus, the frequencies of the parent haplotypes decrease to 90% of their original frequencies. With this set of new haplotypes, we get an average ROH frequency of 16% and an average regional LD of 80.8%. When we repeat the procedure but increase the new haplotype frequencies to 20% of their 'parent' haplotype, ROH frequency dropped to 13% while regional LD remains high at 79%. From this simulation, we observe that ROH detection is very sensitive to heterozygosity present either due to mutation or genotyping errors, whereas the LD in the region is largely preserved despite the mutations introduced. By not allowing any heterozygosity, we miss detecting older ROH in many individuals and this affects the formation of the common regions. So, to capture the LD/haplotype structure using ROHs, it is important to allow a small percentage of heterozygosity.

Confidence scores calculation

For each individual ROH region identified by PennCNV, let n be the number of SNPs within the region. Let x be the number of heterozygous calls among the SNPs within the region, and assume $x \sim Binomial(n; p)$, where p is the unknown probability of observing a heterozygous SNP genotype within an ROH region. Then, the upper bound for p at $\alpha = 0.1$ can be calculated by solving the equation:

$$P(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} = 0.05$$

We use $c = -\log(p)$ as a confidence score; the higher c is, the more likely the called ROH region is a true ROH.

From Figure S1, we see that in general, regions identified on the Affymetrix platform have lower confidence scores as compared to those identified on Illumina platform. Taking only regions whose confidence scores are above the 25th percentile of the confidence scores, Affymetrix platform detects 71522 ROHs of which 57% are also detected on Illumina platform. Illumina detects 46402 ROHs of which 89% are detected on Affymetrix platform. When we increase the confidence cut off to the 75th percentile, Affymetrix detects 7289 ROHs, of which 97% are also detected on Illumina platform. Illumina detects 15432 individual ROHs of which more than half are not detected on Affymetrix platform. This means that majority of Illumina regions at low confidence cut off are detected by Affymetrix, and at high confidence cut off, Illumina platform detects almost all of what Affymetrix platform can detect and about 7000 more ROHs that Affymetrix platform is unable to detect.

Regions not detected by Affymetrix platform may be due to low SNP density in the Affymetrix array in those regions or due to higher noise levels in the intensity data resulting in lower sensitivity in detection. We decide to use Illumina as our platform for detection but still use the combined genotypes from both platforms in the calculation of confidence scores.

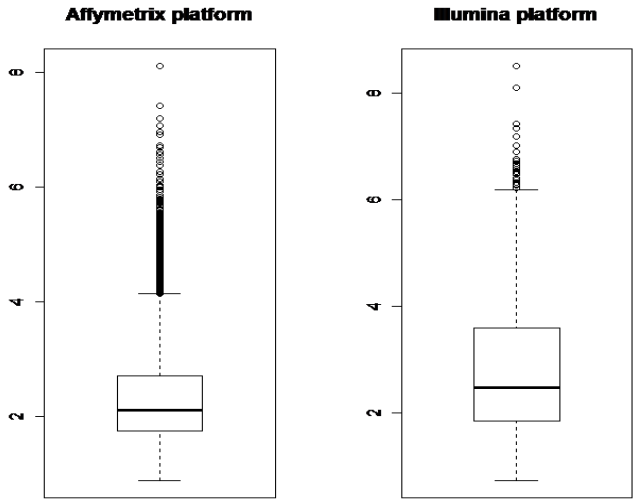


Figure S1: Boxplots of confidence scores of individual ROH regions detected by Affymetrix and Illumina platforms.

Plots of regional LD versus frequency of ROH for Chinese and Indian populations

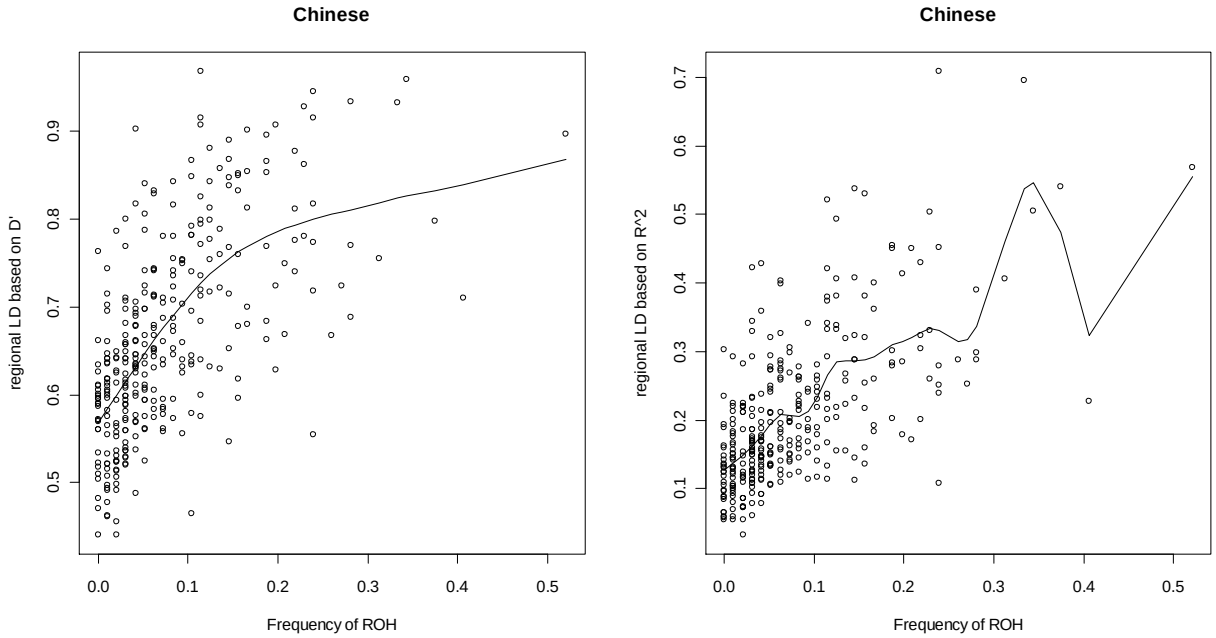


Figure S2: Regional LD vs. frequency of ROH based on a) D' matrix b) R^2 matrix for Chinese population

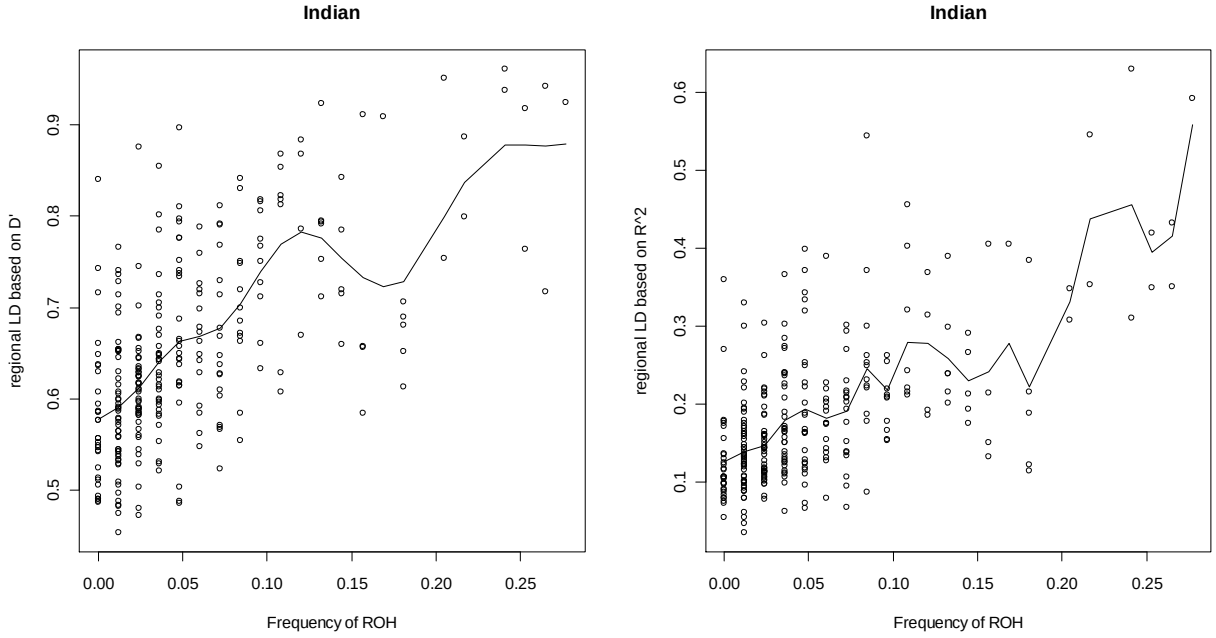


Figure S3: Regional LD vs. frequency of ROH based on a) D' matrix b) R^2 matrix for Indian population

Results based on 414 common regions (when individual regions are filtered at the 75th percentile confidence score cut off)

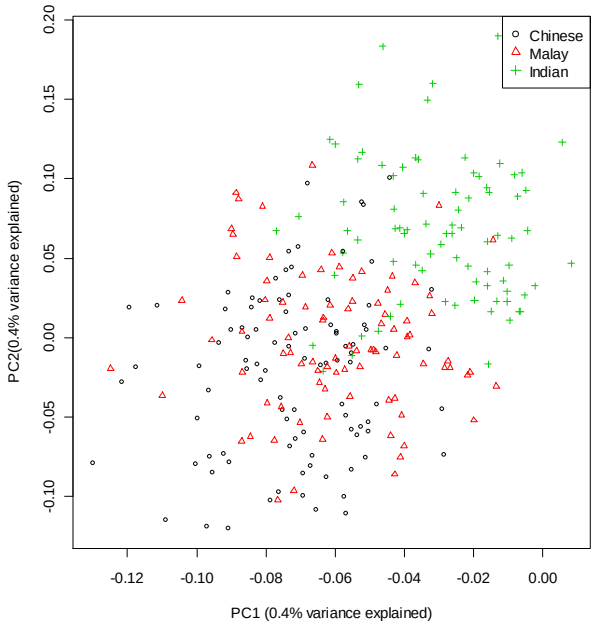


Figure S4: Principal component 2 vs. principal component 1 using absence/presence of 414 common ROHs.

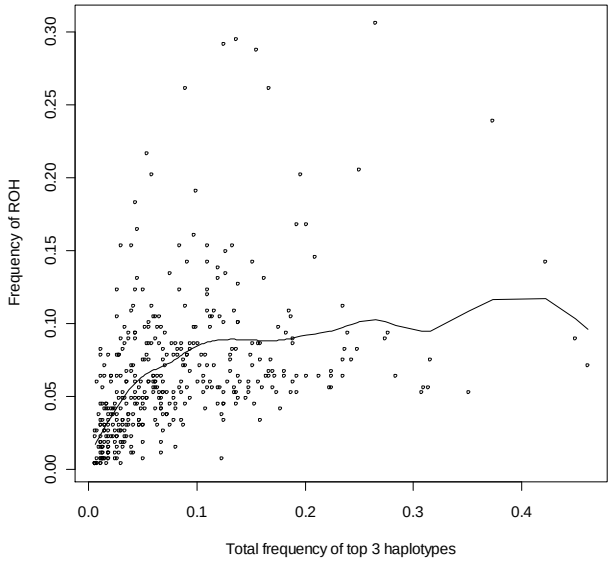


Figure S5: Frequency of ROH loci vs. total frequency of top 3 haplotypes for 414 common ROH loci

Comparison with VarLD (results based on 414 common regions)

Out of 39 regions with differential LD, 11 overlap with a ROH locus, as compared to 29 regions which overlap with a ROH loci when we used a less stringent confidence cut off to identify common ROH regions (see main manuscript).

Comparison with regions associated with positive selection (results based on 414 common regions)

Out of 30 positively selected regions, 10 overlap with a ROH locus, as compared to 28 regions which overlap with a ROH loci when we used a less stringent confidence cut off to identify common ROH regions (see main manuscript).

Table S1: Regions which overlap with 'ROH islands' found in Europeans (from Nothnagel*et al*., 2010's study)

Regions from Nothnagel			Regions from thisstudy					
Chr	Start	End	Start	End	Numberof			Total (%)
					Indians	Chinese	Malays	
14	65754607	66956534	65758913	66794163	20	23	18	21.3
4	33305316	34167260	33400308	34604048	4	7	7	6,3
3	50382348	51835857	50936680	51979357	10	11	8	10.1
12	110249612	111461573	110613553	111448453	21	45	26	32.1
1	35023369	36505444	35201015	36357664	17	2	1	7.0
5	129845818	131423014	129573162	130581277	7	10	5	7.7
11	47998479	49391209	46931701	48858129	3	3	2	2.8
16	65360598	66845475	65280236	66985827	7	16	7	10.5
16	46391563	46826430	46509288	47291485	2	16	4	7.7
10	74211870	75086795	74460386	74990059	22	29	23	25.8

Table S2: Regions which overlap with high frequency ROHs (fromAuton*et al*., 2009's study)

Regions fromAuton				Regions from thisstudy					
Chr	Start (Kb)	End (Kb)	Pop.	Start (Kb)	End (Kb)	Numberof			Total (%)
						Indians	Chinese	Malays	
1	15380	17401	E.Asia	15960	17101	1	1	2	1.4
2	176872	177858	E.Asia	177149	178254	0	3	1	1.4
3	43179	45125	E.Asia	44110	45092	1	8	3	4.2
4	32251	34658	Europe	32994	34443	3	4	3	3.5
4	32251	34826	Mexico	32994	34443	3	4	3	3.5
4	32528	34431	S.Asia	32994	34443	3	4	3	3.5
4	32555	34431	E.Asia	32994	34443	3	4	3	3.5
4	158335	160168	E.Asia	158098	159384	1	1	1	1.0
10	21520	23314	E.Asia	21679	23079	4	2	2	2.8
15	61190	64122	E.Asia	61487	63420	0	1	2	1.0
17	53118	54759	E.Asia	53356	54746	0	2	5	2.4

Table S3: ROH loci with above 30% population frequency

Chr	Start	End	Length	No. of indiv. (%)
3	48656057	49679991	1023935	140 (52.2)
3	49384708	50133778	749071	115 (42.9)
17	55785079	56526387	741309	105 (39.2)
1	28443393	29001138	557746	103 (38.4)
15	62134999	62858424	723426	97 (36.2)
12	98743338	99292958	549621	93 (34.7)
12	110613553	111448453	834901	92 (34.3)
1	8341341	8813712	472372	89 (33.2)
1	50538569	51263192	724624	88 (32.8)
1	73374782	73890423	515642	87 (32.5)

Mean and median pairwise D'/R^2 for SNPs within each ROH region

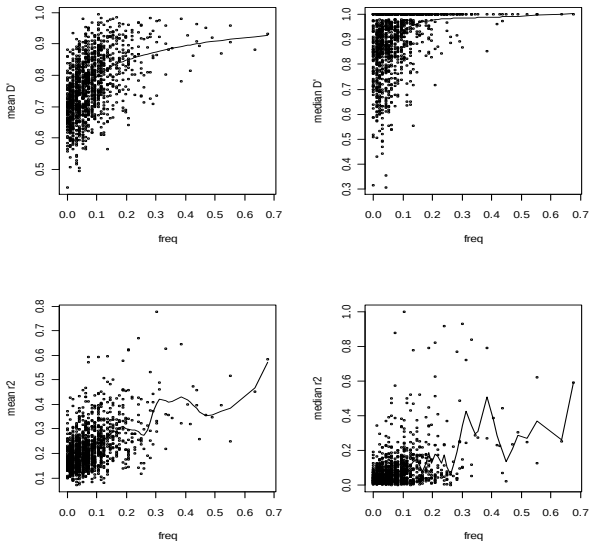


Figure S6: Mean and median pairwise D'/R^2 for SNPs within each ROH region (D'/R^2 calculated up to a distance of 250kb from each SNP). Results shown here are for the Chinese population. Every SNP with minor allele frequency ≥v5% has a chance to bedefined as the focal SNP and for each focal SNP, we compute the LD between the focal SNP and all otherSNPs with MAF ≥N5% that are found within 250kb upstream and downstream of the focal SNP.

PCA using random non ROH regions

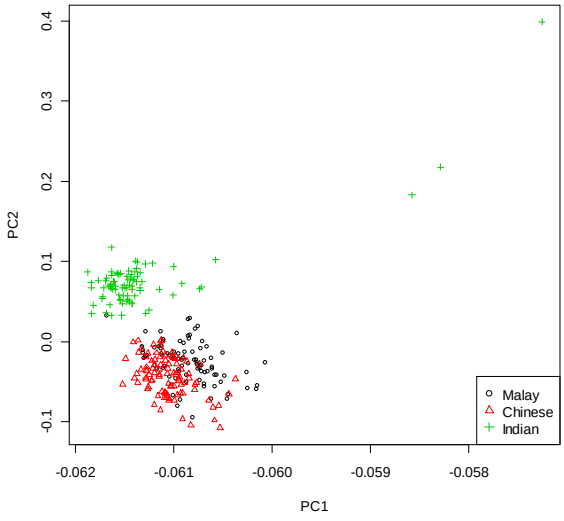


Figure S7: Results of PCA on haplotype frequencies of ROH regions (Based on 1256 random non-ROH regions)

Positive selection

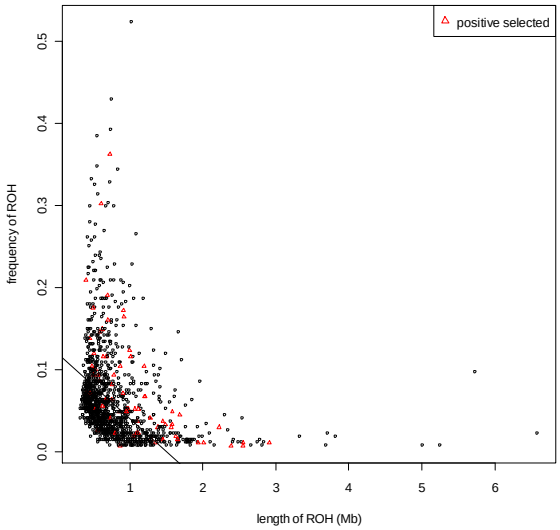


Figure S8: An ROH that has higher frequency than would be expected for its length may also be evidence of positive selection. For example, if we draw an arbitrary line on the plot of frequency of ROH vs. length of ROH, such that points above the line are suggestive of possible candidates of positive selection. 54/544(10%) of the points above the line overlaps with a region found to be positively selected, while only 15/712(2%) of the points below the line overlaps with a region found to be positively selected (p-value = 2.041e-09).

Confidence scores

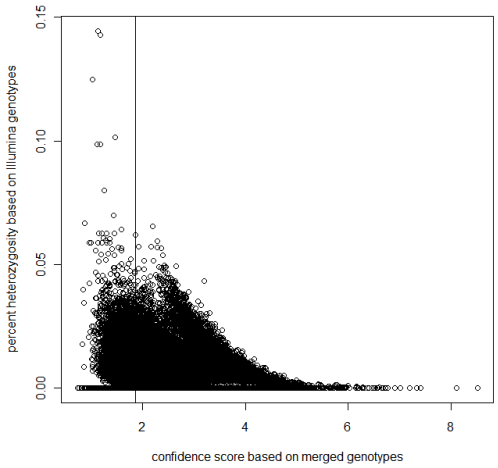


Figure S9: Vertical line drawn is at 25th percentile of the confidence scores. Majority of regions have less than 5% heterozygosity.

Review

Statistical challenges associated with detecting copy number variations with next-generation sequencing

Shu Mei Teo[1,2,3,*], Yudi Pawitan[3], Chee Seng Ku[3], Kee Seng Chia[1,2], and Agus Salim[1,*]

[1]Saw Swee Hock School of Public Health, National University of Singapore, Singapore

[2]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore

[3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Analysing next-generation sequencing (NGS) data for copy number variations (CNVs) detection is a relatively new and challenging field, with no accepted standard protocols or quality-control measures so far. There are by now several algorithms developed for each of the four broad methods for CNV detection using NGS, namely the depth of coverage (DOC), read-pair (RP), split-read (SR) and assembly-based (AS) methods. However, due to the complexity of the genome and the short read lengths from NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs, no matter which method or algorithm is used.

Results: In this review, we describe and discuss areas of potential biases in CNV detection for each of the four methods. In particular, we focus on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulty in identifying duplications. To gain insights to some of the issues discussed, we also download real data from the 1000 Genomes Project and analyse its DOC data. We show examples of how reads in repeated regions can affect CNV detection, demonstrate current GC correction algorithms, investigate sensitivity of DOC algorithm before and after quality-control of reads and discuss reasons for which duplications are harder to detect than deletions.

Contact: g0801862@nus.edu.sg, agus_salim@nuhs.edu.sg

1 INTRODUCTION

Copy number variations (CNVs) are an important and abundant source of variation in the human genome, encompassing a greater proportion of the genome as compared to single nucleotide polymorphisms (SNPs); an estimated 1.2% of a single genome differs from the reference human genome when considering CNVs, as compared to 0.1% by SNPs (Pang *et al.*, 2010). In the last several years, SNP arrays and array comparative hybridization (aCGH) are commonly used for detection of CNVs, albeit with relatively low resolution especially in terms of breakpoint determination. Sanger sequencing of paired reads, often seen as the gold standard for CNV detection, is able to detect CNVs with higher accuracy and resolution, to detect balanced rearrangements such as inversions and translocations, as well as to detect CNVs in regions where probe density of other platforms, such as SNP arrays, is low. However, the technique is not feasible for a large number of genomes due to time and budget constraints. Next-generation sequencing (NGS) or also known as high-throughput sequencing (HTS) attempts to combine the benefits of array technology and sequencing. The biggest advantage of NGS over traditional Sanger sequencing is the ability to sequence millions of reads in a single run at a comparatively inexpensive cost (Metzker, 2010). However, due to the complexity of the genome and the short read lengths (usually 35-400bp) from NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs, no matter which method or algorithm is used.

The growing popularity and success of NGS is evident from large scale projects such as the 1000 Genomes Project (http://www.1000genomes.org/), which aims to sequence at least 1000 individuals from different populations around the world in order to construct a detailed map of genetic variations in the human genome (The 1000 Genomes Project Consortium, 2010). Thus far, in its pilot phase, the project has identified approximately 15 million SNPs, 1 million short indels and more than 20,000 structural variations (SVs), most of which were previously unreported (about 61% of deletions and 89% of duplications are novel). The average SV size detected by the study was 8 kb, approximately four times smaller than a recent SV detection study using tiling CGH array (Conrad *et al*., 2010). SVs include dosage-alterating variants such as CNVs (usually defined as deletions and insertions larger than 1Kb) and shorter indels, as well as dosage-neutral variants such as inversions and translocations.

Nevertheless, current analytical methodologies to analyse NGS data for CNVs are not yet mature and there are no well-established pipelines/protocols/quality-control measures. Broadly, there are four methods for CNV detection using NGS data, namely (1) depth of coverage (DOC, or also known as read-depth (RD) methods), (2) paired-end mapping (PEM, or also known as read pair (RP) methods), (3) split-read (SR) and (4) assembly-based (AS) methods (Alkan *et al.*, 2011). The different methods are usually complementary to one another as the underlying concepts excel at detecting certain types of variants, and a large proportion of discovered variants remain unique to a particular approach (Alkan *et al*., 2011). For example, in the 1000 Genomes Project CNV analysis, the group applied various variations of the four methods with a

*To whom correspondence should be addressed.

total of 36 call sets with vastly varying degrees of false discovery rates (FDR <10% - 89%), as well as notable differences in terms of genomic regions ascertained, CNV size range and breakpoint precision among the different methods (Mills *et al.,* 2011). This review paper highlights and investigates challenges encountered when analysing NGS data for CNVs. In particular, we focus on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulty in identifying duplications. Since the characteristics of CNVs in germline and tumour cells are different, we caution that this review focuses largely on CNVs in the germline, and issues unique to tumour CNVs (also known as copy number alterations or CNAs) are not discussed.

2 THE FOUR CLASSES OF METHODS FOR CNV DETECTION USING NGS

We describe each of the four methods for CNV detection using NGS data, namely (1) depth of coverage (DOC), (2) paired-end mapping (PEM), (3) split-read (SR) and (4) assembly-based (AS) methods. Except for the latter, the other three classes of methods require first mapping the sequenced reads to a known reference genome. We summarize a list of commonly used software for CNV detection using NGS data in Table 1. Readers may refer to seqanswers website: http://seqanswers.com/wiki/Software for a more comprehensive list.

The underlying concept of identifying CNVs using DOC is similar to that of using intensity data: a lower than expected DOC/intensity indicates deletion and a higher than expected DOC/intensity indicates duplication. Most DOC methods count the number of reads that fall in each pre-specified window of a certain size (Abyzov *et al*., 2011; Xie *et al*., 2009; Yoon *et al.*, 2009). The algorithm relies heavily on the assumption that the sequencing process is uniform, i.e. the number of reads mapped to a region is assumed to follow a Poisson distribution and is proportional to the number of copies. However, certain biases such as GC-content and mappability cause this assumption to be unrealistic. Regions of the genome may be over or under-sampled regardless of the copy number of the region, often resulting in spurious signals. Most DOC algorithms correct for GC-content bias before detecting CNVs (Abyzov *et al.,* 2011; Yoon *et al.,* 2009) while there are others that use ratios between reads from the target and reference genome and claims to mitigate the need for GC-correction if the two datasets are prepared in the same way (Xie *et al.,* 2009). Other algorithms also exploit SNP heterozygosity information or also known as 'B allele frequency' to call CNVs and loss of heterozygosity (LOH) regions (Boeva *et al*., 2012; Sathirapongsasuti *et al*., 2011). DOC algorithms usually detect large CNVs and are unable to detect copy neutral events such as inversions and translocations. Single end or paired end data may be used for this analysis.

PEM methods require the reads to be paired (Chen *et al*., 2009; Hormozdiari *et al.,* 2009; Korbel *et al.,* 2009). The concept is that the fragments of DNA from which the reads are to be sequenced have a fragment length (or also known as insert size) of a certain distribution. When the sequenced ends of the fragment map to the reference at a distance longer than expected, it is indicative of a deletion in the studied genome. Vice versa, when the sequenced ends of the fragment map to the reference at a distance shorter than expected, it is indicative of an insertion in the studied genome. Based on the patterns from which the paired reads are mapped to the reference, PEM can also detect inversions and translocations (see Xi *et al*., 2011 for a review of the different SV signatures in PEM). For example, if the two ends of a fragment are mapped with a wrong orientation, it could be an indication of an inversion (Feuk, 2010). The size of CNVs detected using PEM is limited by the insert size and as a result, PEM often detects smaller CNVs.

Split read (SR) methods focus on pairs of reads where one read is mapped uniquely to the reference while the other read failed to be aligned (Ye *et al.,* 2009). The idea is that the location of the unmapped read may span the breakpoint of the CNV. The mapped read is used as an anchor to narrow down the search space for the split read alignment of the unmapped read. SR analysis has the advantage of being able to pinpoint the location of the breakpoints.

Assembly-based (AS) methods, on the other hand, do not align the reads to a known reference but construct the genome piece-by-piece; this is also known as *de novo* sequencing. Some AS methods still use the reference genome as a guide to resolve repeats. This is known as *comparative assembly* (Pop *et al.,* 2004). AS methods can discover new non-reference sequence insertions. AS methods work best for small genomes such as bacterial genomes and are less widely used in NGS sequencing of humans because the short reads from NGS makes assembly in repeat regions difficult (Ye *et al*., 2009). Most AS algorithms for NGS data are extensions of the method described by Pevzner *et al.,* 2001 which uses de Bruijn graphs. It is difficult to judge which method is superior although the methods developed more recently such as SOAPdenovo (Li *et al.,* 2010) claims faster computation time and longer contig size and assembly accuracy when compared to earlier methods such as ABySS (Simpson *et al.,* 2009) and velvet (Zerbino *et al.,* 2008). Cortex (Iqbal *et al.,* 2011) is capable of assembling multiple genomes simultaneously.

Some algorithms use a combination of methods for more accurate detection of CNVs. For example, CNVer (Medvedev *et al.,* 2010), HYDRA (Quinlan *et al.,* 2010) and SVDetect (Zeitouni *et al.,* 2010) supplements DOC with PEM information. Genome STRiP combines information from DOC, PEM, SR as well as other features of sequence data at population level (Handsaker *et al.,* 2011). Genome STRiP is one of the highest performing methods used in the 1000 Genomes pilot Project, indicating that there is benefit in combining different approaches (Mills *et al*., 2011).

Table 1. Commonly used software for CNV detection using NGS data

Program	Reference	Comments
Depth of coverage		
CNVnator*	Abyzov *et al.,* 2011	Uses mean shift approach on fixed window GC-content adjusted read counts.
Rdxplorer*	Yoon *et al.,* 2009	Uses event-wise testing on fixed window GC-content adjusted read counts.
SeqCBS	Shen *et al.,* 2012	Gives approximate confidence intervals for assessing confidence in the segmentation.
CNVseq	Xie *et al.,* 2009	Uses ratios between reads from target and reference genome.
SegSeq	Chiang *et al.,* 2009a	Segments genomes of a tumour and matched normal sample by a sliding fixed size window.

		Boundary is refined after change point is called.
ExomeCNV	Sathirapong-sasuti *et al.*, 2011	For exome sequencing data. Uses read count ratio to detect CNVs, and B allele frequencies to detect LOH
Control-FREEC	Boeva *et al.*, 2012	Uses total coverage and B allele frequencies of SNPs to call CNVs and LOH.
Paired end Mapping		
Variation Hunter*	Hormozdiari *et al.*, 2009	Based on maximum parsimony. Uses soft clustering.
breakdancer*	Chen *et al.*, 2009	Consist of two complementary algorithms: BreakDancerMax predicts insertions, deletions, inversions, inter- and intra-chromosomal translocations, BreakDancerMini predicts small indels.
PEMer*	Korbel et al. 2009	Clusters long and short events separately. Confidence value for each SV. Built in database and simulation program.
Split read		
Pindel*	Ye *et al.*, 2009	Uses pattern growth algorithm Identifies breakpoints of large deletions and medium sized insertions
Assembly based		
Cortex*	Iqbal *et al.*, 2011	capable of assembling multiple genomes simultaneously
SOAP denovo*	Li *et al.*, 2010	claims faster computation time and longer contig size and assembly accuracy when compared to earlier methods such as ABySS and velvet
Velvet	Zerbino *et al.*, 2008	-
ABySS	Simpson *et al.*, 2009	-
Combination/others		
Genome STRiP*	Handsaker *et al.*, 2011	Combines DOC, PEM and distribution of evidence across samples and within a genomic locus
HYDRA	Quinlan *et al.*, 2010	DOC + PEM
ABI tools	McKernan *et al.*, 2009	CBS
Spanner*	Mills *et al.*, 2011	Uses PEM, able to find tandem duplications
SVDetect	Zeitouni *et al.*, 2010	DOC + PEM Competible with SoLiD and Illumina paired-end reads.

*used in 1000 Genomes Project

3 DATA SETS

For the purpose of gaining insights to the issues we are about to discuss, we download sequenced data of individual NA12891 that was deeply sequenced (> 20X coverage) by the Illumina Genome Analyzer platform as part of the 1000 Genomes pilot Project (The 1000 Genomes Consortium, 2010). The reads are paired, 36 bases in length and are aligned to the human reference build 36 (hg18) using the MAQ aligner (Li *et al.*, 2008). The aligned reads are downloaded in BAM format from http://www.1000genomes.org/.

MAQ calculates a phred-scaled quality score for each read/pair of reads that is equal to minus ten times the common logarithm of the probability that a read is wrongly aligned; a quality score of 30 indicates a 1 in 1000 probability that the read is incorrectly mapped. When a read can be mapped equally well to more than one location, a random position is chosen out of all equally possible positions, and the reads are assigned a quality score of zero; these reads are termed *multi-reads* (Harismendy *et al.*, 2009; Treangen *et al.*, 2012). Different aligners have different approaches of dealing with multi-reads. For example, the aligner 'micro-read fast alignment search' (mrFAST) reports all suitable positions of multi-reads (Alkan *et al.*, 2009).

3.1 Estimating depth of coverage

We estimate DOC by counting the number of reads, based on their start positions, in non-overlapping windows of 100 bases. This is the current strategy of most DOC algorithms (Yoon *et al.*, 2009; Abyzov *et al.*, 2011).

3.2 Pre-filtering criteria

For DOC calculation, we keep only reads which are flagged properly aligned, termed 'read mapped in proper pair' in Picard (http://picard.sourceforge.net/explain-flags.html), and reads that are not paired (i.e singletons). About 67% of the reads are flagged properly aligned and about 23% are singletons. We exclude reads which are technical duplicates, paired reads where one read in the pair is unmapped and other reads which are not 'mapped in proper pair'. Singletons are reads where only one end of the fragment is sequenced either due to library preparation or sequencing failure of one the reads in a pair (Alexej Abyzov, personal communication, Nov 2011). It should be differentiated from reads which are paired but where only one of the reads in the pair is mapped to the reference. Singletons are informative and should not be filtered. This is illustrated in Figure 1, which shows obvious signal of decreased DOC using only singleton reads in a region validated to be a deletion by Mills *et al.*, 2011. In the 'Phred score filtered dataset', we further remove 7% of the reads whose mapping quality is less than 30 (but not zero). About 14% of the reads have a mapping quality of zero. These multi-reads are reads that cannot be uniquely aligned to a single position in the genome, meaning that there exists more than one location where the read can be mapped to equally well. We observe the patterns of multi-reads in regions with known CNVs to investigate how these reads can affect CNV detection.

3.3 Reference CNVs

We use the integer copy-numbers for a total of 5,037 CNV loci from Conrad *et al.* (2010)'s study as well as McCarroll *et al.* (2008)'s study as a reference list. Conrad's experiments were done as follows: First, a set of 20 NimbleGen arrays, each comprising 2.1-million oligonucleotide probes were used to generate a new map of CNV locations. Subsequently, a customized Agilent CGH-

platform comprising of 105,000 oligonucleotide probes was used to detect the loci and the genotypes were estimated for 450 HapMap samples using a Bayesian algorithm with stringent selection for optimal normalization and cluster locations for every locus (See Supplementary Methods in Conrad *et al*. 2010 for more details). In total, for individual NA12891, there are 517 deletions (copy-number less than 2) and 253 duplications (copy-number more than 2). It should be noted, however, that a true gold-standard reference list for CNVs is not available, and this list does not have 100% sensitivity and specificity.

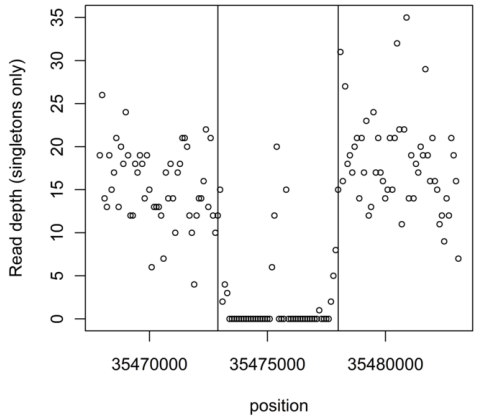


Figure 1: DOC using singletons only (deletion region in Chromosome 22: 35472901 – 35478000). This figure shows that singleton reads independently provide informative evidence of a deletion in this region.

3.4 SNP array intensities

We download SNP array intensities for the Affymetrix 6.0 array for individual NA12891 from the HapMap 3 project raw data database (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). We use the PennCNV algorithm (Wang *et al*., 2007) to obtain Log R ratios (LRR), using samples from the third phase of the HapMap project as the reference panel.

3.5 High-confidence regions

In order to investigate the reasons for the discordance between the reference regions and DOC data, we plot DOC data for specific regions and observe patterns in the data. To narrow down our search for interesting regions, we limit this analysis to high-confidence regions which we define as follows: A deletion region from the reference list is considered 'high-confidence' if it also shows an average LRR of less than log(0.5) ~ -0.7. A duplication region is considered 'high-confidence' if it shows an average LRR of greater than log(1.5) ~ 0.4. There are 60 high-confidence deletions and eight high-confidence duplications. The regions range from 1Kb to 156Kb, and the number of SNP markers range from 1 to 73.

4 REPEAT REGIONS AND MAPPABILITY ISSUES

NGS technology produces mainly short reads, and this poses a challenge in the alignment to the reference genome because reads that fall in repetitive regions in the genome cannot be mapped unambiguously. Furthermore, mutations or sequencing errors in one or two locations may also cause reads to be mapped wrongly (Li *et al.*, 2008). In the 1000 Genomes trios Project, about 20% of the reference genome was considered inaccessible (defined as regions with many ambiguously placed reads or unexpectedly high or low numbers of aligned reads). The resulting low sensitivity in detecting CNVs in repeated/segmental-duplicated regions is a serious problem, because there is an observed enrichment of CNVs in segmental duplicated regions and many breakpoints lie in duplicated regions (Medvedev *et al.*, 2009). This class of CNVs is one of the most poorly studied variants as previous technologies for CNV detection such as aCGH and SNP arrays also have problems resolving them.

For AS methods, repeat regions create challenges because if the read length is shorter than the repeat region, it is not straightforward to decipher the original sequence since overlap between the reads or contigs will be ambiguous (Knudsen *et al.*, 2010). For other methods that require mapping to a reference, there are different alignment strategies for dealing with multi-reads, such as (1) discarding the reads, (2) choosing a position at random out of all equally good match positions, and (3) reporting all possible positions.

The first strategy limits the analysis only to unique regions of the genome, and may miss many CNVs. Moreover, when using DOC methods, excluding multi-reads may cause the identification of false deletions, i.e., regions with a large number of multi-reads will be falsely detected as a deletion if these reads were removed. This is illustrated in Figure 2, which shows a region in Chromosome 20 where several deletions would be falsely identified if multi-reads are excluded. This phenomenon was also observed by Abyzov *et al.* (2011), whose algorithm picked up ten times as many deletions when multi-reads are discarded.

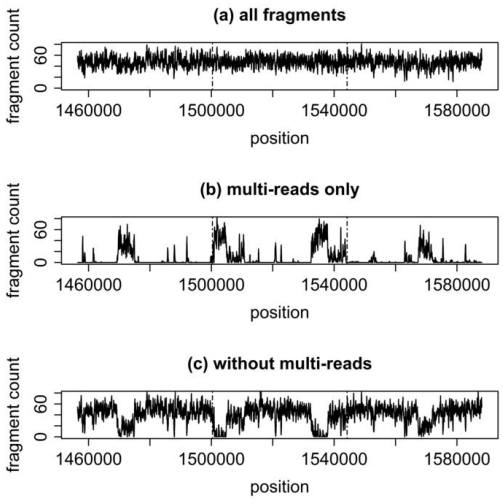


Figure 2: Fragment count for NA12891, Chromosome 20. (a) uses all fragments and shows a relatively flat DOC which varies around the aver-

age. (b) uses only multi-reads and shows several small regions with spikes in multi-reads. (c) uses all fragments with multi-reads removed, we observe "holes" or dips in DOC that would be identified as deletions by DOC algorithms. Multi-reads are placed at random out of all equally possible locations.

Placing a multi-read at random (strategy 2) is also not ideal: For example, a true deletion may exist in a region where there exist similar sequences elsewhere in the genome, causing multi-reads to be mapped to the deletion region where there is supposed to be less or none, thereby diluting the signal (Figure 3). This suggests that the alignment strategy of discarding multi-reads or random placement of multi-reads is inadequate for detecting duplications in repeated regions. A better strategy incorporating other alignment methods and other kinds of data is needed to identify these regions.

He *et al*. (2011) developed a new algorithm for tandem copy number variation reconstruction in repeat-rich regions which considers all locations of possible mappings and uses information on PEM and DOC. Alkan *et al.* (2009) developed a new alignment method, mrFAST; the aligner maps short sequence reads to a repeat-masked reference genome, meaning that all loci with known high-copy common repeats were first masked before alignment, and reports all mapping locations for multi-reads. It also keeps track of mutation in multi-reads. This method has been shown to be able to predict absolute copy number and multicopy differences. Sudmant *et al*. (2010) also uses a similar approach to identify and genotype CNVs within segmental duplications. However, these approaches seem to work only for deeply sequenced data (>20X), and more has to be done to extend these methods for lower coverage data (Chiang *et al.*, 2009).

Longer read lengths from third generation sequencing (TGS) may partially solve the problems with repeats, but even with a read length of 1kb, there still remains about 1.5% of the human genome sequence that is non-unique (Schatz *et al.*, 2010).

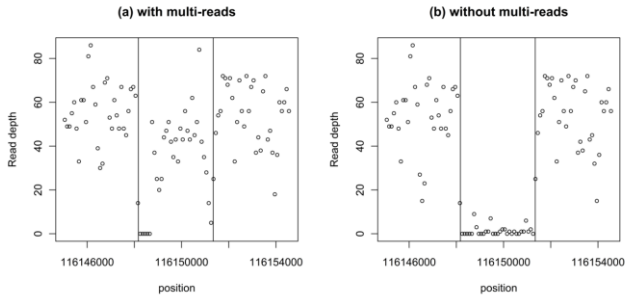


Figure 3: High-confidence deletion region in Chr. 4 (116148170 - 116151343) not identified by DOC methods in Mills *et al.* (2011). (a) Some evidence of deletion is seen when we include all reads. (b) Deletion signal becomes more obvious with multi-reads removed.

5 GC CONTENT

It has been observed that depth of coverage has a unimodal relationship with GC-content (Yoon *et al*., 2009; Abyzov *et al.*, 2011; Benjamini *et al.*, 2012), where regions with high or low GC-content have decreased DOC. Harismendy *et al.* (2009) also observed that unique sequences present at equimolar quantities in library generation end up being sequenced at vastly different DOC. This bias causes problems in all methods. For example, in PEM or SR methods, a region of low depth of coverage may have insufficient reads for enough evidence to discern the variants at that location. For AS methods, regions with low coverage may also result in insufficient information to infer a continuous sequence (Knudsen *et al.*, 2010). The problem can however be solved by increasing the overall sequence depth. The most affected of four methods by GC-content bias is the DOC method.

DOC algorithms rely heavily on the assumption that the sequencing process is uniform, so that the depth of coverage can be assumed to be proportional to the copy number. However, when there are biases that cause sequencing depth to differ for reasons other than the change in copy number, it makes differentiating true deletions/duplication from under/over-sampled regions in the genome difficult. Previous published algorithms correct for GC-content by adjusting the DOC in the window using the GC-content of the window (Yoon *et al*., 2009; Abyzov *et al.*, 2011). This method of correction may be inadequate as the choice of bin size is often arbitrary. Moreover, several studies have observed that it is the GC-content of the full DNA fragments, not only the reads, which cause most of the bias (Benjamini *et al*., 2012).

A recently developed algorithm for GC correction considers the GC-content of the fragment and can produce base-pair resolution predictions of GC-content bias (Benjamini *et al*., 2012). We applied the method on this dataset but observed an increase in overall variance of DOC after correction. Hence, we did not use the results of this correction for subsequent comparisons (see Supplementary Materials for more details).

We download the GC-content per five bases from the UCSC genome bioinformatics website: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/gc5Base/. We correct for GC-content bias in a similar fashion as described by Yoon *et al*. (2009). The GC-corrected DOC was calculated using the following equation:

$$RD^{i}_{corrected} = RD_{global} \times RD^{i}_{raw}/RD_{gc},$$

where i is the bin index, RD_{global} is the average DOC over all bins in the chromosome (we used a trim mean, omitting 5% of bins from both extremes), RD^{i}_{raw} is the DOC for the i^{th} window before correction, and RD_{gc} is the median DOC of all windows with the same GC-content. Since there are few windows with GC less than 20 or greater than 75, for robustness, we set the lower/upper limits for GC in a window to 20 and 75 respectively. Figure 4 (left) plots the DOC of the windows versus the GC percentage of the windows. We observe a similar unimodal relationship between DOC and GC-content as reported by previous papers. In AT rich regions, coverage increases with increasing GC, and in GC rich regions, coverage decreases with increasing GC. The peak coverage can be different for different data sets and different chromosomes but is usually located between 0.35 to 0.5 GC. Figure 4 (right) shows that GC-content bias is removed after correction. However, it is worth noting that even though GC-content bias is removed, the variance in DOC remains rather large, meaning that not all local variations in DOC are associated with GC-content and thus cannot be removed by the GC-correction.

The cause of GC-content bias is speculated to be largely due to PCR amplification step in NGS (Benjamini *et al.*, 2012; Aird *et al*., 2011). Since PCR amplification is not required in TGS, bias ob-

served in DOC due to PCR may be resolved (Schadt *et al*., 2010). The longer read lengths of TGS will also improve challenges caused by the short read lengths of NGS. However, since TGS technology is still new, it is premature to comment on its performance, and too soon to judge if TGS can fulfil its promises of advancement over NGS.

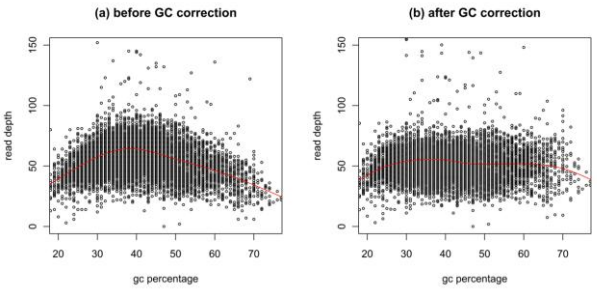


Figure 4: Read depth versus GC percentage before and after correction.

6 PHRED SCORE FILTERING

There has been little documentation on how read mapping quality affects CNV calling. Most algorithms state a default filtering criteria without any substantial evidence for the choice. For example, PEM algorithm, BreakDancer, uses a default filter of mapping quality > 10 (Chen *et al*., 2009) while DOC algorithm, Rdxplorer, uses a default filter of mapping quality > 30 (Yoon *et al.*, 2009).

7 COMPARISONS

We perform sensitivity analysis to investigate the effects of GC-correction and Phred-score filtering. We compare three methods: (1) GC-corrected and Phred-score filtered, (2) GC-corrected but not filtered by Phred-score, and (3) Phred-score filtered but GC-uncorrected. For each CNV in the reference list, we use the t-statistic to determine if the DOC in the region is significantly increased/decreased. For each deletion region *i*, we calculate the *t*-statistic as such:

$$t_i = \frac{\bar{x}_i - c\hat{\mu}_g}{\hat{\sigma}_i / \sqrt{n_i}},$$

where $\bar{x}_i$ is the average DOC in the region, $\hat{\mu}_g$ is the global average DOC for the chromosome where the region lies, $\hat{\sigma}_i$ is the standard deviation of the DOC in region *i*, n_i is the number of windows in the region and *c* is a constant which we vary from 0.5 to 0.8. For each duplication region *j*, the t-statistic is calculated in a similar fashion:

$$t_j = \frac{k\hat{\mu}_g - \bar{x}_j}{\hat{\sigma}_j / \sqrt{n_j}},$$

where *k* varies from 1.2 to 1.6. For each set of tests, we account for multiple comparisons using the False Discovery Rate (FDR). A region is identifiable if the FDR is less than 0.01.

Table 2 shows that there are little differences in sensitivities for all three methods, suggesting that both GC-correction and Phred-score filtering does not seem to be crucial in the sensitivity of detection of CNVs. It should be noted however that this analysis does not investigate the specificity of CNV detection. Overall, GC-correction and Phred-score filtering lowers the variance of DOC, indicating the potential of minimizing the number of false positive regions identified. However, this is a simple and limited analysis and further studies are needed to discern the benefits of GC-correction and filtering by Phred score.

Table 2. Sensitivity of Phred-score filtered and unfiltered datasets, and GC-corrected and non-GC-corrected datasets.

	deletion				duplication				
k	0.5	0.6	0.7	0.8	1.6	1.5	1.4	1.3	1.2
Phred score filtered + GC-corrected	0.31	0.66	0.83	0.89	0.07	0.12	0.15	0.2	0.24
Phred score unfiltered	0.31	0.66	0.82	0.89	0.09	0.12	0.16	0.21	0.25
GC-uncorrected	0.32	0.65	0.81	0.89	0.08	0.11	0.15	0.2	0.25

8 INSERTIONS ARE HARDER TO DETECT THAN DELETIONS

For all methods, identifying duplications has been acknowledged as more challenging as compared to identifying deletions. With regards to PEM methods, the bias against detection of insertions is because PEM detects insertions when the mapped reads are at a distance shorter than the fragment length and hence it is unable to detect insertions larger than the insert size of the reference library, or more specifically the length upper bound of an insertion detected is the average fragment length minus the length of the reads (Wang *et al*., 2008; Hormozdiari *et al*., 2009). This is evident in detection of CNVs using PEM of the diploid Asian 'YH' genome, where 2441 deletions were identified as compared to 33 duplications (Wang *et al*., 2008).

In DOC methods, we observe that the sensitivity of detecting deletions and duplications is around 89% and 25% respectively for the best case scenarios (Table 2). This observation is similar to that observed in Abyzov *et al*. (2011), who estimated that about 90% of deletions identified by aCGH or SNP arrays are DOC-accessible, while only 20-30% of duplications are DOC-accessible. This may be due to the lack of sensitivity of DOC methods in distinguishing a change in number of copies from *N* to *N*+1, especially if *N* is large. For example, suppose a sequence is repeated twice in the reference genome (*N* = 2) at locations A and B, while the studied genome has an additional copy (*N* = 3). Then, assuming an average of 20X coverage, locations A and B would have an average of 60 reads shared among both locations (following strategy of random placement of multi-reads), meaning an average of 30 reads at both A and B, a 50% increase in DOC. If we increase *N* to 5 in the reference and 6 in the studies genome, then each repeated location in the reference would have an average of 120/5 = 24 reads, only 20% more than the average, and likely to be undetectable due to the high variance in DOC.

In the list of high-confidence regions (see section 'High-Confidence Regions'), all 60 deletions can be found in at least one release set from Mills *et al*. (2011), but four of the regions were not detected by DOC methods. When we plotted the read depth in these regions, we observed that two regions have obvious decreased DOC (figure not shown) and should have been detected while the other two were not detected most likely due to the presence of multi-reads diluting the deletion signal (see Figure 3).

On the other hand, all eight duplications are not identified in any of the release sets (see table S1 and Figures S1-8 in Supplementary Materials). This is partly due to the fact that most release sets in Mills *et al.* (2011) focus mainly on deletions, with few sets reporting duplications/insertions. Even then, of the eight regions, only regions 2 and 5 show distinct elevated depth of coverage; these regions have little or no multi-reads. Among the other six regions that do not show obvious increase in DOC, four of them overlap with known segmental duplication regions (segmental duplicated regions as defined in http://humanparalogy.gs.washington.edu/). This is also supported by the presence of multi-reads in these regions; neither keeping nor removing multi-reads result in strong DOC signal of the presence of duplication.

9 DISCUSSION

Next-generation sequencing, with its ability to perform massive parallel sequencing in a single run, is becoming increasingly popular. This brings with it an unprecedented opportunity to sequence many genomes at a relatively inexpensive cost (as compared to using Sanger sequencing). However, with billions of reads generated per individual, the sheer and exponentially increasing amount of data demands for better bioinformatics support and computers with larger storage and higher computing powers. No less important than the production of the data is the information technology infrastructure and bioinformatics team needed to analyse it, with speculations that the costs associated with handling, storing and analysis of the data could be more than the production of the data.

Analysing NGS data for structural variants is a relatively new and challenging field, with no standard protocols or quality-control measures. The four methods of CNV detection are complementary. Comparing DOC, PEM and SR methods used in the 1000 Genomes Project, each approach uniquely discovered 30% to 60% of the CNVs (Abyzov *et al*., 2011). These three methods require first mapping the sequenced reads to a reference genome. Since the mapped reads are used in all downstream analysis, this first step of alignment is extremely crucial. As has been shown in the paper, how the aligner or subsequent algorithm deals with reads in repeat regions is very important for detecting variants that lie in these regions. Currently, the problem of CNV detection in repeated regions is still not completely solved.

Using real data from the 1000 Genomes Project, this paper highlights and investigates challenges associated with current methodologies and areas of potential biases encountered when analysing NGS data for CNVs. In particular, we focus on issues pertaining to (1) mappability, (2) GC-content bias, (3) quality-control measures of reads, and (4) difficulty in identifying duplications. We feel this is a timely critical review that would aid researchers in a much needed well-validated pipeline for the analysis of structural variants.

ACKNOWLEDGEMENTS

This work is supported by the Swedish Science Council and National University of Singapore Start-up Grant No. R-186-000-103-133. SMT acknowledges support from the National University of Singapore Graduate School for Integrative Sciences and Engineering (NGS) Scholarship.

REFERENCES

Abyzov A *et al.* (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research* **21**: 974-984.

Alkan C *et al.* (2011) Genome structural variation discovery and genotyping. *Nature Review Genetics* **12**: 363-376.

Alkan C *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41**: 1061-1067.

Aird D *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**: R18.

Benjamini Y, Speed TP (2012) Summarizing and correction for GC-content bias in high throughput sequencing. *Nucleic Acids Research* DOI:10.1093/nar/gks001.

Boeva V *et al.* (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**: 423-425.

Chen K *et al.* (2009) BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Nature Methods* **6**: 677–681.

Chiang DY *et al.* (2009a) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6**: 99-103.

Chiang DY, McCarroll SA (2009b) Mapping duplicated sequences. *Nature Biotechnology* **27**: 1001-1002.

Conrad DF *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.

Dohm JC *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**: 16.

Feuk L (2010*)* Inversion variants in the human genome: role in disease and genome architecture. *Genome Medicine 2: 11.*

Harismendy O *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**:R32. DOI: 10.1186/gb-2009-10-3-r32.

Handsaker RE *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on population scale. *Nature genetics* **43**: 269-276.

He D *et al.* (2011) Efficient algorithms for tandem copy number variation reconstruction in repeat-rich regions. *Bioinformatics* **27**: 1513-1520.

Hormozdiari F *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research* **19:** 1270-1278.

Iqbal Z *et al.* (2011) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics* **44**: 226-232.

Korbel J *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology* **10**: R23.

Knudsen B *et al.* (2010) A computer simulator for assessing different challenges and strategies of *de novo* sequence assembly. *Genes* **1**: 263-282.

Li H *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**: 1851-1858.

Li R *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**: 265-272.

McCarroll SA *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166-1174.

McKernan *et al.* (2009) Sequence and structural variation in a human genome uncovered by massively parallel ligation sequencing using two-base encoding. *Genome Research* **19**: 1527-1541.

Medvedev P *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods Supplement* **6**:11.

Medvedev P *et al.* (2010) Detecting copy number variation with mated short reads. *Genome Research* **20:** 1613-1622.

Metzker ML (2010) Sequencing technologies – the next generation. *Nature reviews* **11**: 31-46.

Mills RE *et al*. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59-65.

Pang AW *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11**: R52. DOI: 10.1186/gb-2010-11-5-r52.

Pevzner PA *et al.* (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 9748-9753.

Pop M *et al.* (2004) Comparative genome assembly. *Briefings in Bioinformatics* **5:** 237-248.

Quinlan AR *et al.* (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research* **20**: 623-635.

Sathirapongsasuti JF *et al.* (2011) Exome sequencing-based copy number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**: 2648-54.

Schadt EE *et al.* (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**: R227-240.

Schatz MC *et al.* (2010) Assembly of large genomes using second generation sequencing. *Genome research* **20**: 1165-1173.

Shen JJ, Zhang N (2012) Change-point model on nonhomogeneous poisson processes with application in copy number profiling by next-generation DNA sequencing. *The Annals of Applied Statistics* **6**: 476-496.

Simpson JT *et al.* (2009) ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**: 1117-1123.

Sudmant PH *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**: 641-646.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061 - 1073.

Treangen T J, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Review Genetics* **36**: 13.

Wang K *et al.* (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**: 1665-1674.

Wang J *et al. (*2008*)* The diploid genome sequence of an Asian individual. *Nature* **456***:* 60-65.

Xi R *et al.* (2011) Detecting structural variations in the human genome using next generation sequencing. *Briefings in functional genomics* **9**: 405-415.

Xie C, Tammi MT (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**: 80.

Ye K *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865-2871.

Yoon S *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586-1592.

Zeitouni B *et al.* (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* **26**: 1895-1896.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**: 821-829.

Supplementary Materials for
Statistical challenges associated with detecting copy number variations with next-generation sequencing
Shu Mei Teo, Yudi Pawitan, Chee Seng Ku, Kee Seng Chia, Agus Salim

Table S1. List of eight high-confidence duplications

Region	Copy Number	Chr.	Start	End	Overlap segmental duplications?
1	4	3	127155055	127161736	Y
2	3	4	34360003	34365018	N
3	4	5	150860656	150863376	N
4	4	10	81128690	81226767	Y
5	3	10	90784968	90793035	N
6	4	17	41626903	41724649	Y
7	4	17	41568592	41626007	N
8	4	20	1500338	1544222	Y

The following plots S1 – S8 are known duplications for this individual found in previous studies and have evidence of increased LRR in SNP array.

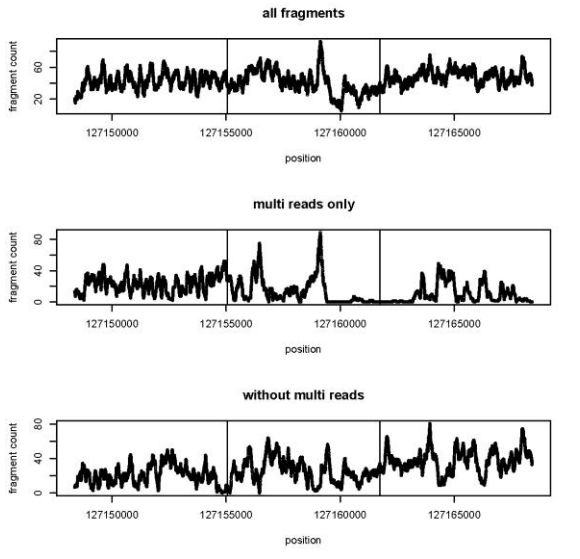


Figure S1: High confidence duplication, region 1 in Chromosome 3: 127155055 – 127161736

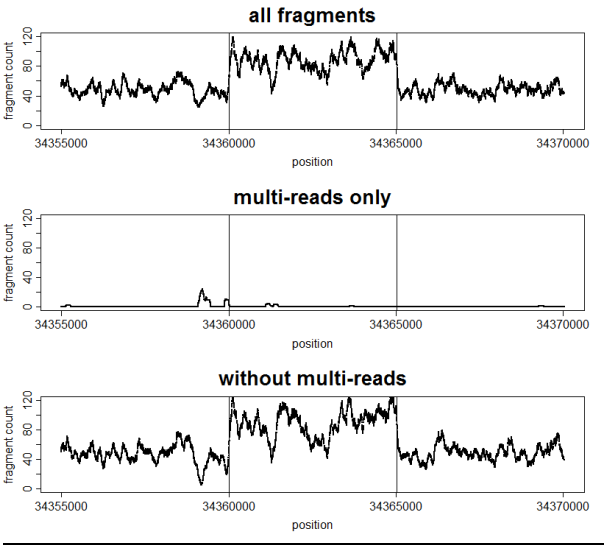


Figure S2: One of the eight high-confidence duplications (Region 2). This example shows a positive duplication signal in the region. This region is not in a repeated region, indicated by the lack of multi-reads (middle panel).

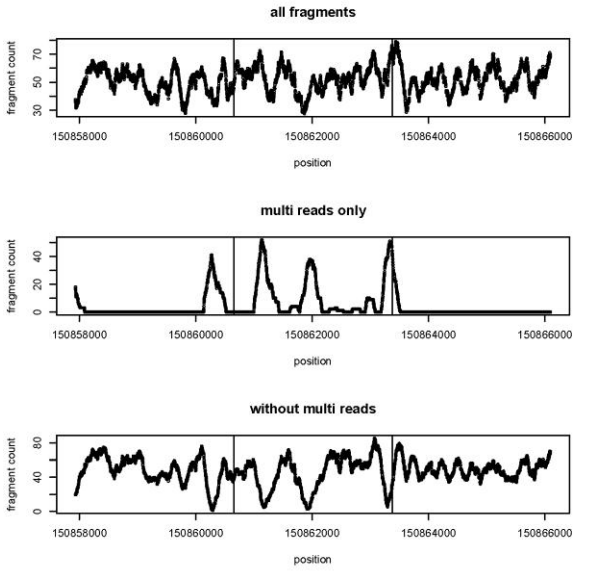


Figure S3: High confidence duplication, region 3 in Chromosome 5: 150860656 – 150863376

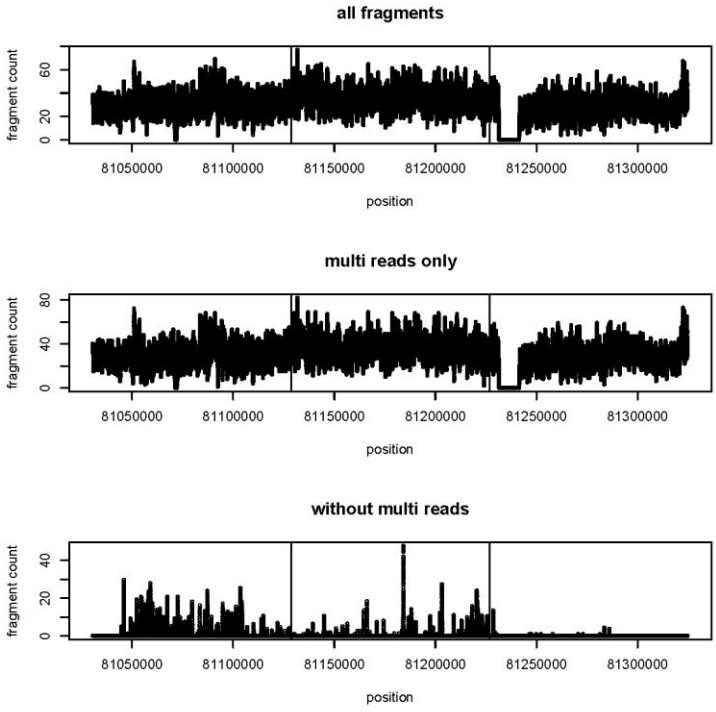


Figure S4: High confidence duplication, region 4 in Chromosome 10: 81128690 – 81226767

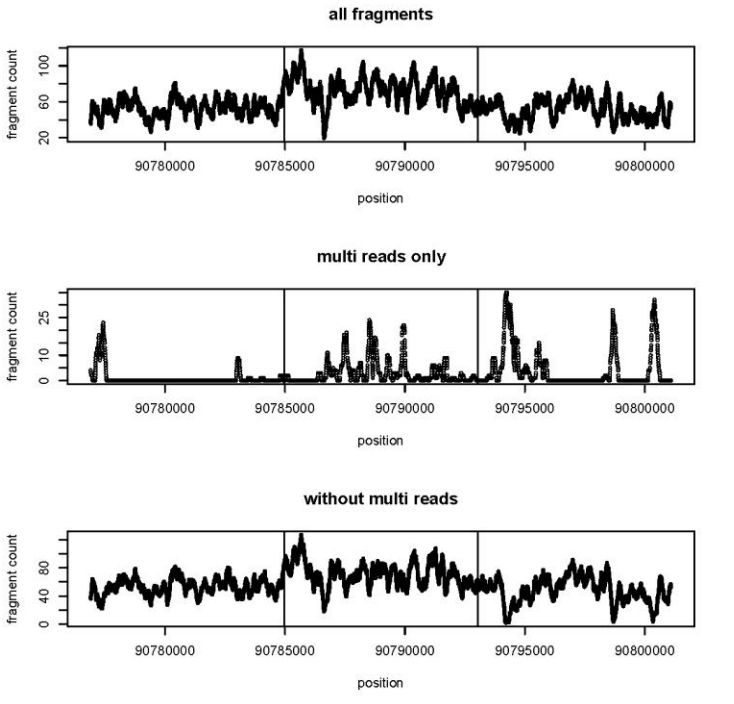


Figure S5: High confidence duplication, region 5 in Chromosome 10: 90784968 – 90793035

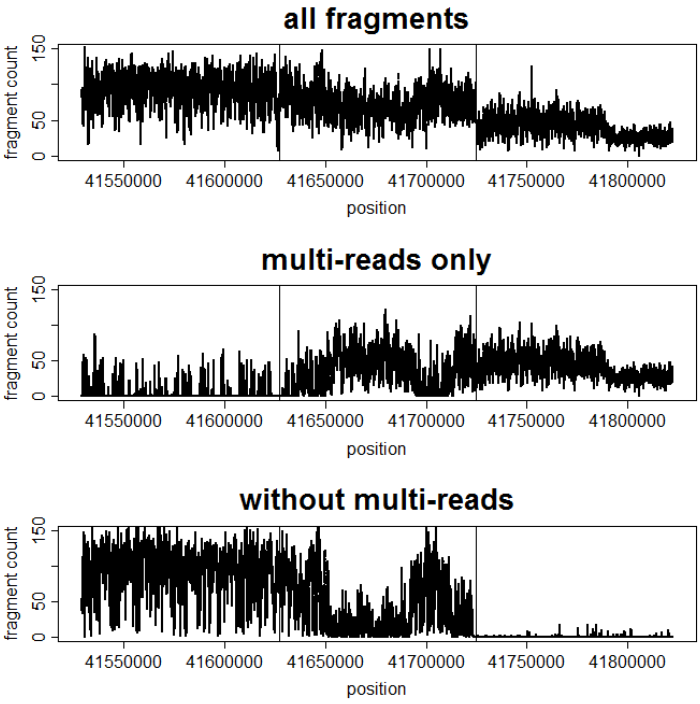


Figure S6: One of the eight high-confidence duplications (Region 6). Top panel shows fragment count using all fragments. Middle panel shows fragment count using only multi-reads. Bottom panel shows fragment count after removing multi-reads.

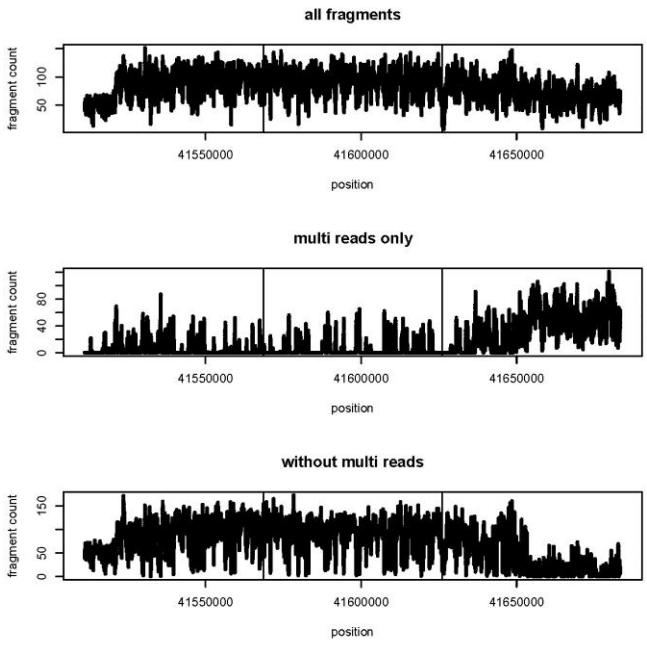


Figure S7: High confidence duplication, region 7 in Chromosome 17: 41568592 – 41626007

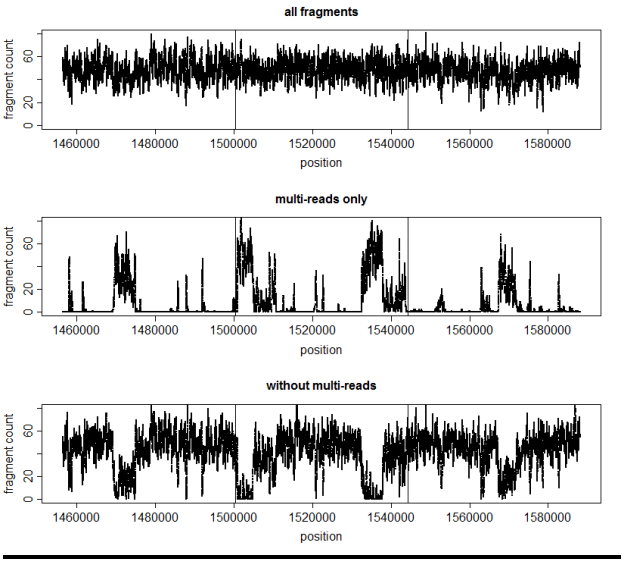


Figure S8: High confidence duplication, region 8 in Chromosome 20: 1500338 – 1544222

GC correction (Benjamini *et al.*, 2012)

The package 'GCcorrect' which implements the GC correction method described in Benjamini et al 2012 requires raw sequence reads for input into the aligner 'Bowtie'. This step is required to generate a list indicating which positions in the reference are not mappable. Since, for this paper, our starting data contains reads that have been already aligned using MAQ, we skip this step and generate the list of unmappable positions as downloaded from http://wiki.rglab.org/index.php?title=Public:Mappability_Profile

All other procedures follow defaults in the package. Using Chromosome 22 as an example, we observe a similar GC effect (as in our GC-correction analysis) using the default plotting parameters of read counts in 10kb windows (Figure S9). However, at the optimal bin size of 599-bases bins, based on largest TV score as suggested in the paper, the correction did not improve overall variance (Figure S10).

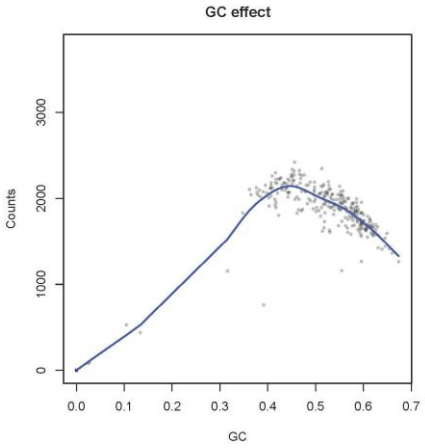


Figure S9: GC curves (10 kb bins) in Chromosome 22.

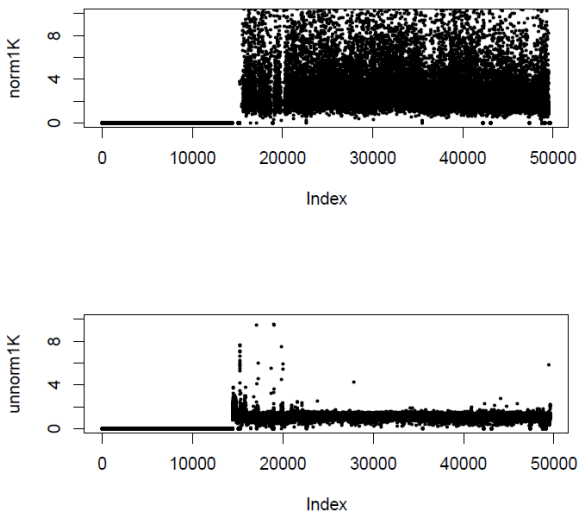


Figure S10: Normalized and un-normalized read depth.

Journal of Human Genetics (2011) 56, 524–533

www.nature.com/jhg

ORIGINAL ARTICLE

A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals

Shu-Mei Teo[1,2,3,5], Chee-Seng Ku[2,5], Nasheen Naidoo[2], Per Hall[1], Kee-Seng Chia[1,2,4], Agus Salim[4] and Yudi Pawitan[1]

The abundance of copy number variants (CNVs) and regions of homozygosity (ROHs) have been well documented in previous studies. In addition, their roles in complex diseases and traits have since been increasingly appreciated. However, only a limited amount of CNV and ROH data is currently available for the Swedish population. We conducted a population-based study to detect and characterize CNVs and ROHs in 87 randomly selected healthy Swedish individuals using the Affymetrix SNP Array 6.0. More than 600 CNV loci were detected in the population using two different CNV-detection algorithms (PennCNV and Birdsuite). A total of 196 loci were consistently identified by both algorithms, suggesting their reliability. Numerous disease-associated and pharmacogenetics-related genes were found to be overlapping with common CNV loci such as CFHR1/R3, LCE3B/3C, UGT2B17 and GSTT1. Correlation analysis between copy number polymorphisms (CNPs) and genome-wide association studies-identified single-nucleotide polymorphisms also indicates the potential roles of several CNPs as causal variants for diseases and traits such as body mass index, Crohn's disease and multiple sclerosis. In addition, we also identified a total of 14 815 ROHs ⩾500 kb or 2814 ROHs ⩾1M in the Swedish individuals with an average of 170 and 32 regions detected per individual respectively. Approximately 141 Mb or 4.92% of the genome is homozygous in each individual of the Swedish population. This is the first population-based study to investigate the population characteristics of CNVs and ROHs in the Swedish population. This study found many CNV loci that warrant further investigation, and also highlighted the abundance and importance of investigating ROHs for their associations with complex diseases and traits.
Journal of Human Genetics (2011) **56,** 524–533; doi:10.1038/jhg.2011.52; published online 2 June 2011

Keywords: Affymetrix SNP Array 6.0; Birdsuite; copy number variants; PennCNV; regions of homozygosity; Swedish population

INTRODUCTION

There is a growing body of copy number variant (CNV) maps covering different world populations.[1–5] Most of these newer studies used high-resolution methods for detecting CNVs, such as the Affymetrix SNP Array 6.0, which has a higher density of single-nucleotide polymorphism (SNP) and copy number probes than previous microarray-based methods. This has led to an improved performance of microarray-based methods to detect smaller CNVs (<50 kb).[1,6] In contrast, previous studies have used much lower resolution arrays, such as the bacterial artificial chromosome (BAC) clone or oligonucleotide comparative genomic hybridization arrays and SNP genotyping arrays.[7–10] Currently, there is only one CNV-detection study in a Swedish population,[10] but this was performed in a small sample size of 33 individuals and used a low-resolution 32-K bacterial artificial chromosome clone microarray. This has hampered the study from detecting less common and smaller CNVs and from estimating the population frequency of CNVs. The ability to detect smaller CNVs is critical as they are more numerous than the larger CNVs.[11]

In addition, the study by Díaz de Ståhl *et al.*[10] was unable to detect regions of homozygosity (ROHs) as the bacterial artificial chromosome clone microarray was unable to generate allelic intensity data. Research on ROHs has started to gain impetus, as evidenced by the increasing number of publications after the first study by Gibson *et al.*[12] reported the abundance of ROHs in the human genome of outbred populations. Further studies have investigated the population characteristics of ROHs in healthy individuals,[13–15] and also performed association analyses to identify ROHs that are associated with complex diseases and traits in a case–control study design.[16–18]

To circumvent the limitations of the previous study by Díaz de Ståhl *et al.*,[10] we conducted a study in a Swedish population by genotyping 100 individuals using the Affymetrix SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA). The main aim of this study was to perform a more comprehensive detection of CNVs and ROHs in the Swedish

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore and [4]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[5]Joint first author.
Correspondence: C-S Ku, Center for Molecular Epidemiology, National University of Singapore, Singapore 117597, Singapore.
E-mail: csikcs@nus.edu.sg or Professor Y Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden.
E-mail: Yudi.Pawitan@ki.se
Received 18 January 2011; revised 12 April 2011; accepted 25 April 2011; published online 2 June 2011

population and to describe their population characteristics. Although several studies have been performed to detect and characterize CNVs and ROHs in multiple European populations, these studies have also documented the genetic differences among these populations.[14,15,19] The extension of the International HapMap Project to include an additional seven populations in Phase III further suggests that multiple populations from diverse ancestries or different geographical locations are needed to study their population genetics.[20] These previous studies have justified the need for a population-based study to characterize CNVs and ROHs in healthy Swedish individuals. We also compared the Swedish population with the HapMap phase III populations using principal component analysis.

MATERIALS AND METHODS

Samples and genotyping platform

A total of 100 randomly selected healthy Swedish individuals volunteering as controls in case–control studies were studied. Peripheral blood samples of the participants for genomic DNA extraction were drawn and stored at the Karolinska Biobank. Identities of the participants were kept anonymous and no personal identifiers were used. All 100 samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 as per the manufacturer's protocol. Two samples were removed from further analysis because their genotype call rates were below 98% and the remaining 98 samples were used for CNV detection.

CNV-detection algorithms and analyses

CNV calling using PennCNV. We used two CNV-detection algorithms, namely PennCNV[21] and Birdsuite,[22] for both comparison and validation. This study focused only on the CNVs in the 22 autosomes because of the inaccuracy of Birdsuite to detect CNVs in sex chromosomes. Log *R* ratio and B allele frequency were calculated according to the PennCNV algorithm (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affygw6.html). Smaller CNVs (<1 kb) were also included in our analysis, as PennCNV by default does not limit its detection to CNVs >1 kb in size. We applied a set of filtering criteria as recommended by the algorithm, namely Log *R* ratio-s.d >0.35, B allele frequency-median >0.55, B allele frequency-median <0.45 and B allele frequency-drift >0.006 to exclude samples with poor quality of signal intensity data (http://www.openbioinformatics.org/penncnv/). This resulted in a further exclusion of 11 samples, with the final set for analysis consisting of 87 samples. For each sample, PennCNV generated a list of CNVs with their confidence scores. The confidence score is a log Bayes factor that measures the likelihood that the locus harbors an abnormal copy number. A confidence score of $\geqslant 10$ has been recommended as the threshold to classify reliable CNVs. Therefore, we retained all CNVs called with confidence scores $\geqslant 10$ for subsequent analyses. Although the confidence score is only a statistical measure of a true positive, our previous study[5] found that CNVs with a higher confidence score are more likely to be detected consistently across two genotyping platforms. Therefore, this justifies our decision to retain only reliable CNVs called with a sufficient degree of confidence.

Construction of CNV loci using PennCNV output. The CNVs called by PennCNV were shown to overlap across samples. Thus, we merged or grouped these individual CNV calls into discrete, non-overlapping loci, with the boundaries of each locus determined by the union of all CNVs that belonged to that particular locus. This construction of CNV loci was needed to estimate the population frequencies and these steps were performed using the methods that we have developed previously.[5,23] We classified the status of these CNV loci into three categories, 'del' (loci containing deletions), 'dup' (loci containing duplications) and 'del/dup' (loci containing both deletions and duplications).

Copy number polymorphism (CNP) calling using Canary (Birdsuite). Birdsuite software[22] was also used to analyze the Affymetrix SNP Array 6.0 data. There are two components in the software for detecting copy number changes, namely Canary and Birdseye. Canary was used to determine the integer copy number at each of the predefined 1316 CNPs. The term 'CNPs' used by McCarroll *et al.*[1] is to describe common CNV loci. These 1316 CNPs were found in more than one HapMap II individual and their sizes were also accurately determined. Therefore, we used the Canary component in Birdsuite to determine the integer copy number of the 1316 CNPs in the 87 Swedish samples. These 1316 CNPs are distributed in all the autosomes and sex (X and Y) chromosomes. However, 25 CNPs located in the sex chromosomes were removed because the CNP calling in these chromosomes was less accurate. Thus, the results reported in this study comprised only 1291 CNPs in the 22 autosomes. Confidence statistics generated for the CNPs were also used to identify poor-quality calls, and only integer copy numbers detected with high confidence as recommended by the software (confidence score >0.1) were used for subsequent analyses.

Correlation analysis of CNPs. We performed a correlation analysis of CNPs and the nearby SNPs. Because the sizes of the CNPs were previously accurately determined by McCarroll *et al.*,[1] we restricted the analysis to only the CNPs detected by Canary. For each of the 1291 CNPs, SNPs within a 200-kb window from the start and end positions of the CNP were considered. We used the squared Pearson's correlation (r^2) for correlation analysis. The genotype calling of the Affymetrix SNP Array 6.0 was carried out using Birdsuite. In addition, to investigate the potential associations of CNPs with human diseases and traits, the same methods of r^2 calculations for the 1291 autosomal CNPs and the SNPs that were identified by genome-wide association studies (GWAS) were adopted. The list of GWAS-SNPs was downloaded from the National Human Genome Research Institute website (http://www.genome.gov/gwastudies/) on 26 October 2010.

CNV calling using Birdseye (Birdsuite). In addition to PennCNV, we also used another algorithm, Birdseye, to analyze the same set of data as different algorithms tend to have different sensitivities and specificities for detection of CNVs in different regions throughout the genome. As such, CNV loci detected by PennCNV and Birdseye can be cross-validated. Therefore, we used the Birdseye component in Birdsuite to detect additional CNVs throughout the genome, which was not restricted to the 1316 predefined CNPs. Similarly, only CNVs in autosomal chromosomes were used because of the inaccuracy of Birdseye in the sex chromosomes. CNVs with low confidence, as recommended by the software (confidence score $\leqslant 5$), were removed from subsequent analysis.

Construction of CNV loci using Birdseye output. We also constructed CNV loci based on the Birdseye output using methods similar to those applied to the PennCNV output. The cutoff for the confidence score used by PennCNV ($\geqslant 10$) and Birdseye ($\geqslant 5$) was recommended by both algorithms. This allowed for greater comparability between the CNV loci detected by these two algorithms.

Comparison of CNV loci detected by PennCNV and Birdsuite. The CNV loci identified by PennCNV and Birdseye were compared as a 'validation' step. We used a 'reciprocal 50% overlapping' method to compare the CNV loci detected by these two algorithms and considered a CNV locus 'found' by both algorithms when this locus was detected in both PennCNV and Birdseye with an overlap of $\geqslant 50\%$ of their lengths.

Novel CNV loci. To identify novel CNV loci, we compared the CNV loci detected by PennCNV and Birdseye with the data from the Database of Genomic Variants (DGV).[24] We used the latest data from the DGV (variation.hg18.v8.aug.2009.txt and indel.hg18.v8.aug.2009.txt) downloaded from the DGV Website (http://projects.tcag.ca/variation/). A CNV locus identified by PennCNV and Birdseye was considered novel if it did not share at least 50% of its length with any CNV loci cataloged in the DGV. All the downstream analyses after PennCNV and Birdsuite were performed using the statistical software package R (http://www.r-project.org/).

Comparison with HapMap phase III populations

The CEL files of the Affymetrix SNP Array 6.0 for the seven populations in HapMap phase III project were downloaded from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). The HapMap phase III populations studied are people of African ancestry in the southwestern USA (ASW), the Chinese community in Metropolitan Denver, Colorado, USA (CHD), Gujarati Indians in Houston, Texas, USA (GIH), the Luhya in Webuye,

Kenya (LWK), people of Mexican ancestry in Los Angeles, California, USA (MEX), the Maasai in Kinyawa, Kenya (MKK) and the Tuscans in Italy (TSI). All the samples were analyzed using Canary similarly to the analysis of the Swedish population. Only unrelated samples were included in our study, that is, family-related samples were removed using the 'relationships' file provided by the International HapMap Project. After the sample exclusion step, a total of 594 unrelated samples from the seven HapMap III populations were analyzed: ASW (n=52), CHD (n=89), GIH (n=89), LWK (n=90), MEX (n=53), MKK (n=132) and TSI (n=89). We performed principal component analysis to compare the Swedish population with the HapMap phase III populations using the CNP output generated by Canary.

ROH-detection algorithms and analyses

In addition to CNVs, we also detected ROHs using PennCNV in the 22 autosomes of the 87 Swedish individuals. However, we only focused on ROHs $\geqslant$500 kb, as this cutoff was adopted in a previous study.[18] For each of these we confirmed that they are ROHs by determining the genotypes of the SNPs that fall within each region. We then calculated the percentage of heterozygosity (number of heterozygotes/total number of heterozygotes and homozygotes). We also calculated the percentage of missingness genotypes (number of missingness/total number of SNPs in each ROH). First, we used an arbitrary cutoff of the median of the percentage of heterozygosity (2.5%) to allow for some heterozygote calls resulting from calling or genotyping errors. As a result, we removed half of the ROHs with a percentage $>$2.5%. Second, we removed ROHs with $>$1% for the missingness, to remove regions where genotype calling was problematic. Finally, for the remaining ROHs, we also ensured a density of one SNP per 10 kb to exclude those ROHs that could be spuriously detected by a sparse number of SNPs. As such, for a 500-kb ROH, a minimum of 50 SNPs is required. These three criteria were used as the filters to exclude less reliable ROHs. Several summary statistics were then computed to describe the characteristics of ROHs in the Swedish population.

RESULTS

Characteristics of CNVs identified by PennCNV

After filtering unreliable CNV calls, an average of approximately 36 CNVs per individual with a ratio of deletions to duplications of approximately 2.6:1 was discovered (Supplementary Table 1). The number of CNVs per individual ranged from 22 to 65. The median size of a CNV was 28.6 kb and approximately 66% of the CNVs were $<$50 kb and 26% were $<$10 kb (Supplementary Figure 1). The median size of deletions was approximately fourfold smaller than the median size of duplications.

Characteristics of CNV loci identified by PennCNV

We merged overlapping CNVs to construct CNV loci and identified 623 loci, of which 476 loci contained deletions ('del-loci'), 102 loci contained duplications ('dup-loci') and 45 loci contained both deletions and duplications ('del/dup-loci'; Table 1). These 623 loci covered approximately 61.52 Mb of the nucleotide sequence and the sum of the lengths for del-loci (19.83 Mb) was smaller than that for dup-loci (25.80 Mb). Similarly for the individual CNVs (Supplementary Table 1), the average size of del-loci (41.66 kb) was much smaller than that of dup-loci (252.93 kb; Table 1). More than 77% of the del-loci were $<$50 kb, and in comparison only 22.55% of dup-loci were within this size range. The majority (62.75%) of dup-loci ranged from 50 to 500 kb. In summary, there were far more del-loci, but their sizes tended to be smaller than those of dup-loci. A list of the 623 loci is shown in Supplementary Table 2.

Of the 623 CNV loci, 268 loci were detected in $\geqslant$2 individuals (Table 1). The remaining loci were detected in only one individual; these loci were not necessarily 'singleton loci' as we only studied

Table 1 Summary statistics of CNV loci constructed from PennCNV output

Summary statistics of CNV loci (PennCNV output)	*Total*	*Del*	*Dup*
Number of CNV loci	623	476 (76.40%)[a]	102 (16.37%)[a]
Number of CNV loci detected in $\geqslant$2 individuals	268 (43.02%)[b]	194 (40.76%)[b]	29 (28.43%)[b]
Sum of the length of loci (Mb)	61.52	19.83	25.80
Average length per locus (kb)	98.75	41.66	252.93
Average number of markers per locus	58	34	141
Size distribution			
$<$10 kb	141 (22.63%)	132 (27.73%)	6 (5.88%)
$\geqslant$10–$<$50 kb	265 (42.54%)	236 (49.58%)	17 (16.67%)
$\geqslant$50–$<$100 kb	79 (12.68%)	54 (11.34%)	21 (20.59%)
$\geqslant$100–$<$500 kb	110 (17.66%)	52 (10.92%)	43 (42.16%)
$\geqslant$500 kb	28 (4.49%)	2 (0.42%)	15 (14.71%)
Overlapping with DGV			
CNV loci that overlap	388 (62.28%)	298 (62.61%)	54 (52.94%)
CNV loci that did not overlap	235 (37.72%)	178 (37.39%)	48 (47.06%)
Overlapping with UCSC genes			
CNV loci that overlap	202 (32.42%)	135 (28.36%)	51 (50.00%)
CNV loci that did not overlap	421 (67.58%)	341 (71.64%)	51 (50.00%)
Overlapping with CNV loci from Birdseye data and consistent in CNV status that is, del/dup/del+dup			
CNV loci that overlap	196 (31.46%)	160 (33.61%)	30 (29.41%)
CNV loci that did not overlap	427 (68.54%)	316 (66.39%)	72 (70.59%)

Abbreviations: CNV, copy number variant; DGV, database of genomic variants; UCSC, University of California Santa Cruz genes.
[a]The percentage was calculated by dividing 623 loci.
[b]The percentage was calculated by dividing 623, 476 and 102 loci, respectively.
Note: As there are only 45 CNV loci (7.22%) with status del+dup, the summary statistics of these loci are not shown in the table. A full colour version of this Table is available at the Journal of Human Genetics Journal online.

87 individuals. The proportion of del-loci detected in $\geqslant 2$ individuals (40.76%) was much higher than the proportion for dup-loci (28.43%). Among the high-frequency CNV loci (loci that were detected in multiple individuals), several overlapped with disease-related genes such as *WWOX* and *ERBB4* (gastric and pancreatic cancers and melanoma)[25–27] and *CACNA1C* (bipolar disorder)[28] or drug-metabolizing genes such as *GSTT1*[29] (Supplementary Table 2). For example, a deletion locus overlapping with *WWOX* (a tumor suppressor gene) was detected in 24 of the 87 individuals (27.6%), and a deletion locus encompassing *GSTT1* was deleted at a population frequency of 13.8%. In addition, the proportion of del-loci encompassing the UCSC genes (28.36%) was much lower than dup-loci (50.00%) overall.

Detection of CNVs using microarrays is usually plagued with poor specificity or a high false-positive rate. In an effort to validate the 623 CNV loci constructed from the PennCNV output, we compared them with the CNV loci detected by Birdseye. We found 196 loci (31.46%) with $\geqslant 50\%$ reciprocal overlap with the Birdseye data and the status of 'del', 'dup' and 'del/dup' of the 196 loci were consistent with the Birdseye data. For the remaining 427 CNV loci that were not confirmed by Birdseye data, we found that 247 loci had been cataloged in the DGV (please see Materials and methods). Therefore, by applying two different ways of validation, 443 (71.1%) of the 623 CNV loci detected by PennCNV were considered reliable in this study (Table 1).

Characteristics of CNPs identified by Canary (Birdsuite)

Approximately 49.81% of the 1291 autosomal CNPs were non-polymorphic in the Swedish population (Supplementary Table 3). The population frequency distribution pattern of the 1291 CNPs is shown in Supplementary Figure 2. Among the polymorphic loci (648 CNPs) and non-polymorphic CNPs (643 loci) in the Swedish population, 289 loci (44.60%) and 255 loci (39.66%) overlapped with genes or entries from the UCSC annotation of the human genome, respectively. No substantial difference was observed between the polymorphic and non-polymorphic loci.

The majority of the 648 polymorphic CNPs were biallelic (545 CNPs or 84.1%), of which the integer copy numbers were either exclusively deletions, that is, copy number of 0 or 1 (387 CNPs or 59.7%), or exclusively duplications, that is, copy number of 3 or 4 (158 CNPs or 24.4%). Among the biallelic 545 CNPs, only one showed significant deviation from HWE at an FDR <0.01.

Numerous CNPs were found to overlap with important known disease- or pharmacogenetics-related genes (Table 2). The frequencies of these CNPs ranged from relatively uncommon (2.78% for CNP118) to completely polymorphic (100% for CNP88). For example, CNP88 overlapped with GSTM1 and GSTM2 was found to be completely deleted in the Swedish population, where all except one carried two-copy deletions. However, it is noteworthy that in approximately half of the sample (47 individuals), the integer copy numbers were successfully determined with high confidence scores. In addition, high deletion frequencies were also found for CNPs overlapping with other GST enzymes such as GSTT1 (60.00%), GSTT2, GSTT2B and GSTTP1 (98.65%). Two-copy deletion was common for these enzymes—17.6% of the individuals for GSTT1 (CNP2560) and 43.2% for the other GST enzymes (CNP2559).

Besides these phase II metabolizing enzymes, several disease-associated genes were also found to overlap with these CNPs, such as the FCG receptor genes (autoimmune or inflammatory diseases),[30] *TP63*[31] and *WWOX*[26] (lung adenocarcinoma, gastric, pancreatic and other cancers), *CFHR3* and *CFHR1* (age-related macular degeneration),[32] *UGT2B17* (prostate cancer and graft-versus-host disease),[33,34]

Table 2 CNPs that overlap with important and known disease- and pharmacogenetics-related genes

CNP ID	*CN=0*	*CN=1*	*CN=2*	*CN=3*	*CN=4*	*Frequency*	*Chromosome*	*Start*	*End*	*Length*	*UCSC gene (disease/trait)*
118	0	1	70	0	1	2.78	1	159 778 034	159 906 183	128 149	FCGR3A, FCGR2B, FCGR2C, FCGR3B (autoimmune or inflammatory diseases)
11164	0	1	83	2	0	3.49	6	162 658 558	162 660 430	1872	PARK2, parkin (Parkinson's disease)
530	1	10	71	0	0	13.41	3	190 846 372	190 847 332	960	TP63 (cancers)
147	3	31	53	0	0	39.08	1	194 997 658	195 068 695	71 037	CFHR3, CFHR1 (age-related macular degeneration)
603	8	33	46	0	0	47.13	4	69 043 083	69 168 574	125 491	UGT2B17 (prostate cancer, graft-versus-host disease)
2560	15	36	34	0	0	60.00	22	22 680 529	22 726 814	46 285	GSTT1 (phase II metabolizing enzyme)
2203	20	46	17	1	0	79.76	16	76 929 941	76 942 266	12 325	WWOX (cancers)
109	33	39	15	0	0	82.76	1	150 822 330	150 853 218	30 888	LCE3C, LCE3B (psoriasis)
2559	32	41	1	0	0	98.65	22	22 613 016	22 670 785	57 769	GSTT2, GSTT2B, GSTTP1 (phase II metabolizing enzyme)
88	46	1	0	0	0	100.00	1	110 025 907	110 044 476	18 569	GSTM2,GSTM1 (phase II metabolizing enzyme)

Abbreviations: CNPs, copy number polymorphisms; UCSC, University of California Santa Cruz genes.
A full colour version of this Table is available at the Journal of Human Genetics Journal online.

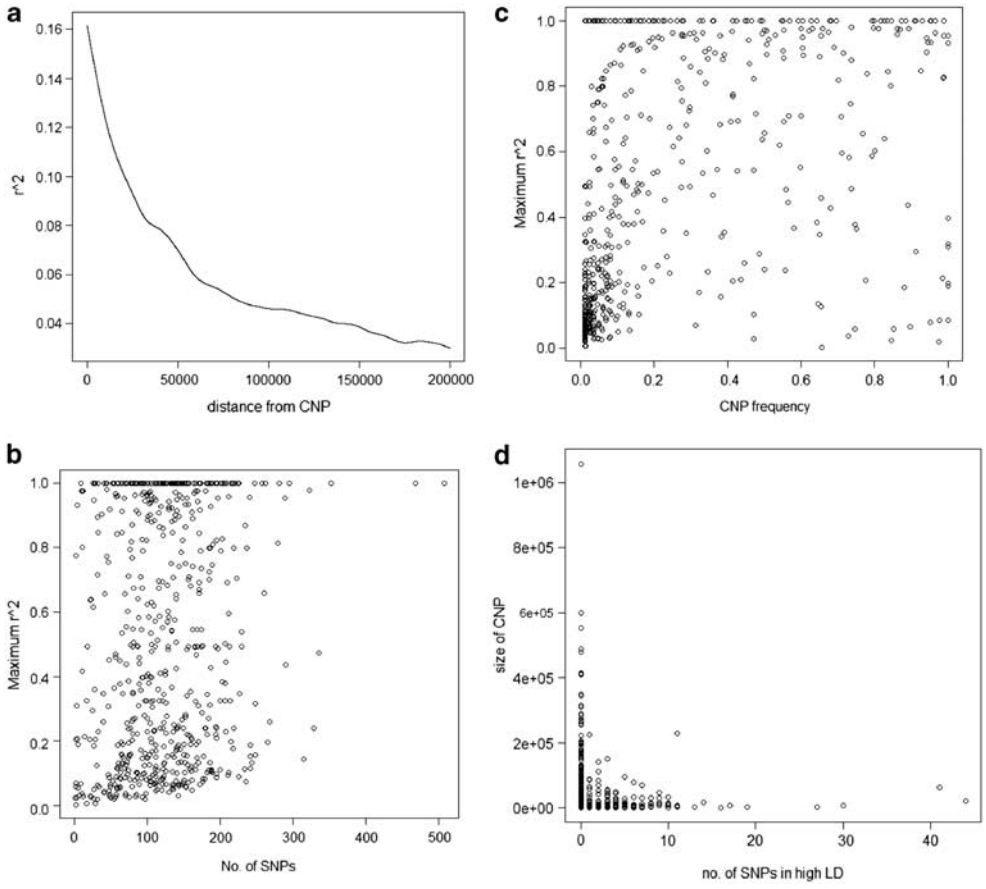


Figure 1 (**a**) The correlation between the r^2 and the distance between copy number polymorphism (CNP) and single-nucleotide polymorphism (SNP). (**b**) Maximum r^2 of CNP versus number of nearby SNPs in 200-kb windows. (**c**) Maximum r^2 of CNP versus CNP frequency. (**d**) Number of SNPs in strong correlation with the size of CNPs.

and *LCE3C* and *LCE3B* (psoriasis and rheumatoid arthritis) among others.[35,36] The high deletion frequency of loci overlapping with *LCE3C* and *LCE3B* (82.76%), *UGT2B17* (47.13%) and *WWOX* (79.76%) requires further studies to investigate their associations with complex diseases such as psoriasis, rheumatoid arthritis and graft-versus-host disease for hematopoietic stem cell transplantation patients. For example, the mismatch of the copy numbers of UGT2B17 was found to be associated with graft-versus-host disease in patients with hematopoietic stem cell transplantation.[34] Deletion of UGT2B17 was also associated with an increased risk for prostate cancer.[33]

Correlation analyses between CNPs and nearby SNPs

To study the correlation patterns with SNPs, we calculated the r^2 between the 648 polymorphic CNPs and nearby SNPs within a 200-kb window from the start and end positions of the CNP. The proportion of the CNPs with at least one SNP in strong correlation ($r^2 > 0.8$) was 31.9%, that is, 207 CNPs were found to be in strong correlation with at least one SNP. The median and maximum numbers of SNPs that were in strong correlation with the 207 CNPs were 3 and 44, respectively. This suggests that half of the 207 CNPs can be tagged by more than three SNPs and some of the CNPs were tagged by tens of SNPs. These results suggest that the majority of CNPs were not being well tagged by the nearby SNPs in the Affymetrix SNP Array 6.0. The strength of the r^2 value decreases with distance between the CNP and SNP (Figure 1a). We further investigated whether CNPs that were not well tagged tend to be located in the genomic regions where SNP markers are sparse. The correlation patterns do not appear to be affected by the number of nearby SNPs and the frequencies of CNPs (Figures 1b and c). In other words, there was no apparent difference in the number of nearby SNPs and the frequencies of CNPs between (a) the CNPs that were in strong correlation ($r^2 > 0.8$) and (b) CNPs that were not in strong correlation with SNPs (Figures 1b and c). However, smaller-sized CNPs were generally in strong correlation with more SNPs than the larger CNPs (Figure 1d).

Correlation analyses between CNPs and GWAS-SNPs

To investigate the potential role of CNPs in the etiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs on the NHGRI GWAS Catalog (http://www.genome.gov/gwastudies/). Of the >3000 GWAS-SNPs that have been found to be associated with various complex diseases and traits, only eight GWAS-SNPs were found to be in strong correlation with six CNPs (Table 3). Following the methods of Conrad *et al.*,[2] we define in our analysis a strong correlation as $r^2 > 0.5$. These eight SNPs were reported to be associated with five diseases or traits, namely body mass index, childhood acute lymphoblastic leukemia, early-onset myocardial infarction, Crohn's disease and multiple sclerosis. Several SNPs were in strong correlation with a single CNP, for example, three SNPs (rs13361189, rs1000113 and rs11747270) were found to be in strong correlation with CNP874.

The most notable SNP was rs2815752 near the *NEGR1* gene (associated with body mass index), which was in perfect correlation ($r^2 = 1$) with CNP60. This locus is a 42-kb deletion located in

Table 3 Correlation between CNPs and GWAS-SNPs at $r^2 > 0.5$

CNP ID	*Chromosome*	*Start position*	*End position*	*Length*	*GWAS-SNP*	*r^2 value*	*Gene*	*Complex disease/trait*
60	1	72 541 504	72 583 736	42 232	rs2815752	1	NEGR1	BMI
147	1	194 997 658	195 068 695	71 037	rs6428370	0.647399825	Intergenic	Acute lymphoblastic leukemia (childhood)
333	2	203 608 045	203 610 291	2246	rs6725887	0.84632626	WDR12	Myocardial infarction (early onset)
874	5	150 185 693	150 198 797	13 104	rs13361189	0.927251567	IRGM	Crohn's disease
874	5	150 185 693	150 198 797	13 104	rs1000113	0.927251567	IRGM	Crohn's disease
874	5	150 185 693	150 198 797	13 104	rs11747270	0.927251567	IRGM	Crohn's disease
877	5	155 409 350	155 415 307	5957	rs4704970	1	SGCD	Multiple sclerosis
933	6	32 539 530	326 81 749	142 219	rs3129934	0.664781909	HLA-DRB1	Multiple sclerosis

Abbreviations: BMI, body mass index; CNPs, copy number polymorphisms; GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.
A full colour version of this Table is available at the *Journal of Human Genetics* Journal online.

chromosome 1 that did not overlap with any of the UCSC genes and is located only 1.3 kb away from the SNP. The total deletion frequency in the Swedish population was high (Table 3 and Supplementary Table 4), of which 51.72% were one-copy deletions and 29.89% were two-copy deletions. CNP874 was found to be in nearly perfect correlation (r^2=0.93) with three GWAS-SNPs located near the *IRGM* gene, which is associated with Crohn's disease. However, in comparison with CNP60, the total deletion frequency for CNP874 was much lower, with only 11.90% one-copy deletions and 1.19% two-copy deletions. This locus spans 13 kb in chromosome 5 and does not overlap with any of the UCSC genes. The three GWAS-SNPs were located 4.8 kb (rs13361189), 21.4 kb (rs1000113) and 40.2 kb (rs11747270) away from the deletion. The CNP877 locus is implicated in multiple sclerosis, where it is in perfect correlation with the GWAS-SNP (rs4704970). None of the individuals were deleted in both copies, and 32.56% were one-copy deletions. The other CNPs were implicated in childhood acute lymphoblastic leukemia (CNP147) and early-onset myocardial infarction (CNP333). Interestingly, all the CNPs found to be in strong correlation with GWAS-SNPs had only deletions in the loci.

Characteristics of CNV loci identified by Birdseye (Birdsuite)

Similar to the PennCNV output analysis, we also merged overlapping CNVs to construct CNV loci for the Birdseye data and identified 641 loci, of which 451 were del-loci, 102 were dup-loci and the remaining 31 were del/dup-loci (Table 4). The proportion of del-loci (76.40%) identified by PennCNV data was higher than that for the Birdseye data (70.36%). In comparison, the Birdseye data identified a higher proportion of dup-loci (24.80%) than the PennCNV data (16.37%). However, these differences are not substantial.

The 641 loci identified by the Birdseye data cover approximately 35.23 Mb of the nucleotide sequence, and the sum of the length for del-loci (13.10 Mb) is smaller than that for dup-loci (15.06 Mb). Similar to PennCNV data, the average size of del-loci (29.04 kb) is much smaller than that of the dup-loci (94.70 kb). However, substantial differences were observed for these parameters between the PennCNV and Birdseye data (Tables 1 and 4). For example, the sum of lengths covering CNV loci detected by the PennCNV data (61.52 Mb) was approximately twice that for the Birdseye data (35.23 Mb), while they have an almost similar number of CNV loci.

More than 60% of del-loci were $<$10 kb, and in comparison, only 18.24% of dup-loci fall within this size range. The majority (52.20%) of dup-loci ranged from 10 to 100 kb. In summary, there were more del-loci, but their sizes tended to be smaller than those of the dup-loci. This is in agreement with the PennCNV data. However, the size distribution pattern of the CNV loci for the Birdseye data is skewed towards the 'smaller' end compared with the PennCNV data. This is apparent when comparing the proportions in the first two strata: (a) $<$10 kb and (b) $\geqslant$10–$<$50 kb between the two sets of data (Tables 1 and 4). The list of the 641 loci is shown in Supplementary Table 2.

Of the 641 CNV loci, 280 loci were detected in $\geqslant$2 individuals (Table 4), and the remaining loci in only one individual. The proportion of del-loci detected in $\geqslant$2 individuals (43.90%) was much higher than the proportion for dup-loci (32.08%). Among the high-frequency CNV loci (loci detected in multiple individuals), several overlapped with disease-associated or pharmacogenetics-related genes such as *WWOX* and *GSTT1*, which have also been observed in the PennCNV data (Supplementary Table 2). Furthermore, the deletion frequencies were comparable between the Birdseye and PennCNV data. For example, a deletion locus overlapped with *WWOX* was also found in the Birdseye data. It was detected in 29 of the 87 individuals (33.33%), and a deletion locus encompassing *GSTT1* was deleted at a population frequency of 11.49%. Among the 196 CNV loci (160 del-loci, 30 dup-loci and 6 del/dup-loci) that were detected by both the Birdseye and PennCNV data and consistent in their CNV status, only 21 loci differed significantly (FDR $<$0.01) in their frequencies estimated by both sets of data. In addition, the proportion of del-loci encompassing UCSC genes (24.83%) was much lower than dup-loci (45.28%); this finding is again consistent with the PennCNV data.

For the CNV loci detected with the Birdseye data, we also performed the 'validation' steps for overlap with the PennCNV data and the DGV. As mentioned earlier, we found 196 loci with $\geqslant$50% reciprocal overlap between the Birdseye and PennCNV data. For the remaining 445 CNV loci that were not confirmed by PennCNV data, we found that 322 loci have been cataloged in the DGV (please see Materials and methods). Therefore, by applying two different ways of validation, 518 (80.81%) of the 641 CNV loci detected by Birdseye were considered reliable in this study (Table 4).

Comparison with HapMap phase III populations

The principal component analysis showed distinct clusters for populations with different ancestries. The first two principal components (PC1 and PC2) separated the African (ASW, MKK and LWK) and non-African (CHD, GIH, MEX, SWED and TSI) populations (Figure 2a). This suggests that the CNP profiles of the African populations were substantially different from those of the non-African populations. From the second and fourth principal components (PC2 and PC4), three distinct clusters were observed (Figure 2b). The three African populations remained as a distinct cluster; however, CHD was separated from the European populations (MEX, SWED and TSI) and the Gujarati Indians (GIH). This indicates that the CNP profile of Gujarati Indians in Houston (Texas, USA) resembles that of the European populations. Principal component analysis was also performed by restricting only the 'European cluster' populations

Table 4 Summary statistics of CNV loci constructed from Birdseye (Birdsuite) output

Summary statistics of CNV loci (Birdseye output)	*Total*	*Del*	*Dup*
Number of CNV loci	641	451 (70.36%)[a]	159 (24.80%)[a]
Number of CNV loci detected in ⩾2 individuals	280 (43.68%)[b]	198 (43.90%)[b]	51 (32.08%)[b]
Sum of the length of loci	35.23 Mb	13.10 Mb	15.06 Mb
Average length per locus	54.96 kb	29.04 kb	94.70 kb
Average number of markers per locus	30	22	42
Size distribution			
<10 kb	303 (47.27%)	272 (60.31%)	29 (18.24%)
⩾10–<50 kb	193 (30.11%)	119 (26.39%)	63 (39.62%)
⩾50–<100 kb	52 (8.11%)	27 (5.99%)	20 (12.58%)
⩾100–<500 kb	79 (12.32%)	31 (6.87%)	40 (25.16%)
⩾500 kb	14 (2.18%)	2 (0.44%)	7 (4.40%)
Overlapping with DGV			
CNV loci that overlap	465 (72.54%)	335 (74.28%)	106 (66.67%)
CNV loci that did not overlap	176 (27.46%)	116 (25.72%)	53 (33.33%)
Overlapping with UCSC genes			
CNV loci that overlap	202 (31.51%)	112 (24.83%)	72 (45.28%)
CNV loci that did not overlap	439 (68.49%)	339 (75.17%)	87 (54.72%)
Overlapping with CNV loci constructed from Birdseye and consistent in CNV status, that is, del/dup/del+dup			
CNV loci that overlap	196 (30.58%)	160 (35.48%)	30 (18.87%)
CNV loci that did not overlap	445 (69.42%)	291 (64.52%)	129 (81.13%)

Abbreviations: CNV, copy number variant; DGV, database of genomic variants; UCSC, University of California Santa Cruz genes.
[a]The percentage was calculated by dividing 641 loci.
[b]The percentage was calculated by dividing 641, 451 and 159 loci, respectively.
Note: as there are only 31 CNV loci (4.84%) with status del+dup, the summary statistics of these loci were not shown in the table.
A full colour version of this Table is available at the Journal of Human Genetics Journal online.

(GIH, MEX, SWED and TSI) in PC2 versus PC4 (Figure 2b). More interestingly, we also found that the CNP profile of the Swedish population was substantially different from that of the other populations such as GIH and MEX, but it was also appreciably different from that of TSI (Figure 2c). These differences further justify the need to detect and characterize the CNV/CNP profile of the Swedish population.

Characteristics of ROHs

By restricting ROHs to ⩾500 kb, a total of 14 815 regions were found in the 87 Swedish individuals with an average of 170 ROHs (Supplementary Table 5). The number of ROHs ranged from 105 to 220. The majority of these ROHs were <1 Mb in length (Supplementary Figure 3). However, by restricting ROHs to ⩾1 Mb, 2814 ROHs with an average of 32 ROHs per individual were found. The median size of the ROHs was approximately 686 kb, with the largest ROH spanning a length of approximately 25 Mb in chromosome 11. This ROH contained 9034 homozygotes, 29 heterozygotes and 2 missing genotypes, and had a density of 3.6 SNPs per 10 kb. The second largest ROH was 12 Mb in length and was detected in a different individual. This ROH contained 1571 homozygotes and 19 heterozygotes and had a density of 1.3 SNPs per 10 kb. The sum of the length of ROHs in each individual (that is, the total length of all the ROHs in one individual) was then computed. It ranged from approximately 87 to 179 Mb with a median and mean of approximately 141 Mb, respectively. This finding suggests that, on average, 141 Mb or 4.92% of the human genome (2867 Mb) was homozygous in these Swedish individuals (Table 5).

The distribution pattern of these ROHs in the 22 autosomes was also studied. The larger chromosomes (chromosomes 1–8) tended to have a higher average number of ROHs per individual (Table 5). For example, these chromosomes had an average number of >9 ROHs per individual, and in contrast, an average number of <5 ROHs per individual was detected in chromosomes 16–22. As a result, chromosomes 1–8 also had a higher average sum of length of ROHs per individual (>7 Mb) than the smaller chromosomes, that is, <4 Mb for chromosomes 16–22. However, this pattern was less obvious when the parameters were adjusted for the sizes of the chromosomes. For example, the proportion of the chromosome encompassed by ROHs for the largest chromosome 1 (4.78%) was smaller than that for the other chromosomes such as chromosome 17 (5.14%). An apparent trend is not observed for the proportion of the chromosome encompassed by ROHs across the 22 autosomes. However, chromosomes 3, 4, 8 and 12 tended to have the highest proportions (5.90–6.16%), and, in contrast, chromosomes 16, 19, 21 and 22 had the lowest proportions (1.76–2.59%). These results were not due to differences in the density of SNPs across the 22 autosomes, as we found no substantial differences in the density of SNPs across the chromosomes (except for chromosome 19, which had a density of <2 SNPs per 10 kb when compared with the other chromosomes). Although chromosomes 3 and 4 had >6% of the proportion of the chromosome encompassed by ROHs, the density of SNPs of these chromosomes was similar to that of chromosome 16, where only approximately 2% of this chromosome was covered by ROHs (Table 5).

DISCUSSION

In this study, >600 CNV loci were detected in the Swedish population using two different CNV-detection algorithms, that is, PennCNV (623 loci) and Birdsuite (641 loci). From these, 196 loci were consistently identified by both algorithms, suggesting their reliability. In addition,

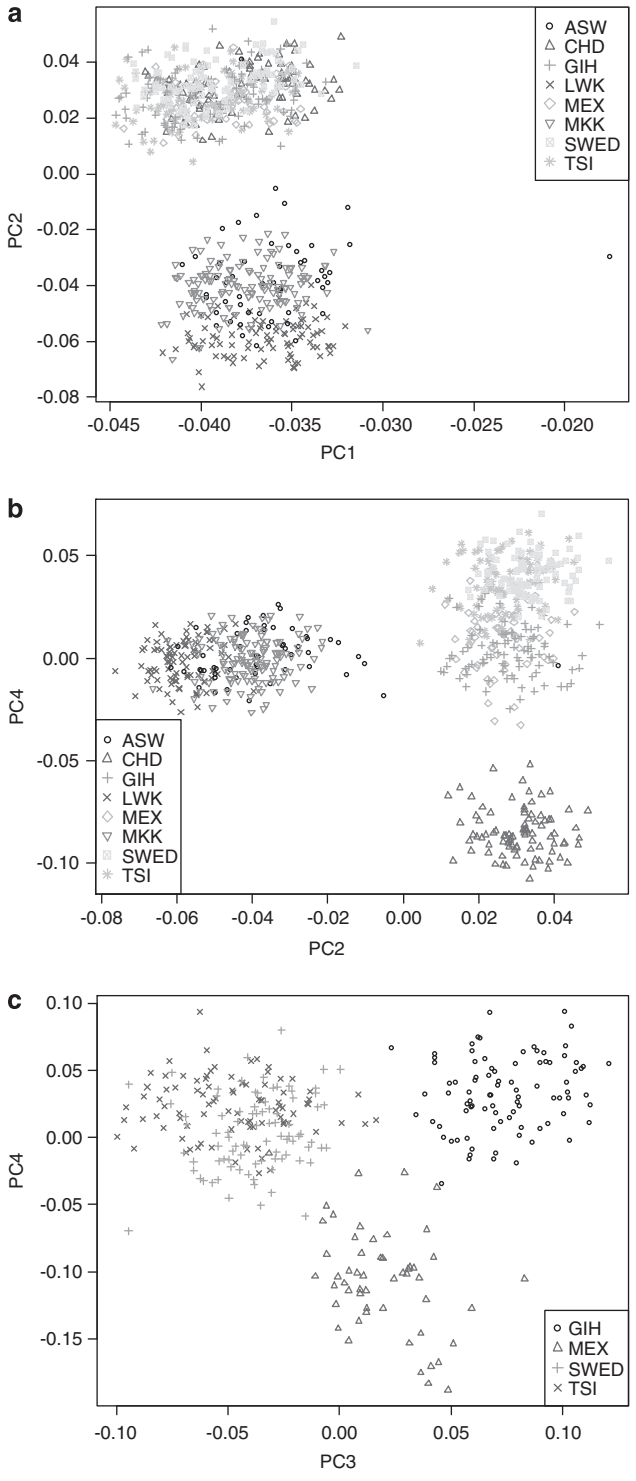


Figure 2 Principal component analysis comparing the populations. (**a**) Swedish and HapMap phase III populations—PC 1 versus PC 2. (**b**) Swedish and HapMap phase III populations—PC 2 versus PC 4. (**c**) Swedish and three HapMap III populations (GIH, MEX and TSI)—PC 3 versus PC 4.

we also identified a total of 14 815 ROHs $\geqslant$500 kb or 2814 ROHs $\geqslant$1 Mb in the Swedish individuals with an average of 170 and 32 regions detected per individual, respectively.

CNVs have been increasingly recognized as a significant source of genetic variation or diversity in human populations. Detection of CNVs using SNP genotyping arrays is more cost-effective and affordable for population-based studies as compared with sequencing-based methods, which are limited to only a few individuals.[37–39] This has enabled our study to investigate the population characteristics of CNVs. Although >600 CNV loci were identified, only 268 were detected in at least two individuals by PennCNV. Similarly, Birdseye also found 280 common CNV loci in the 87 Swedish individuals. More importantly, these common CNV loci were found to encompass several disease-related and important drug-metabolizing genes, suggesting that these loci warrant further characterization and study for their associations with the relevant diseases or traits.

We applied two different algorithms to detect CNV loci as a validation step; 196 loci were found by both the algorithms and these loci were also consistent in their CNV status ('del', 'dup' or 'del+dup'). In the majority of the 196 loci, the population frequencies were also in good agreement between PennCNV and Birdseye data, indicating that these CNV loci are highly reliable. In addition, most of the CNV loci detected by PennCNV (>70%) and Birdseye (>80%) can be 'validated' by comparing them with each other and with the DGV. The proportion of CNV loci overlapping with the DGV was approximately 62% and 72% for PennCNV and Birdsuite, respectively. These percentages could be overestimated because of the false-positive entries in the DGV. Of the 196 CNV loci that were identified by both algorithms, 53 loci had not been previously cataloged in the DGV, which represents a subset of reliable novel CNV loci identified in our study. The list of CNV loci in the DGV is not as yet complete as results from only 42 published studies were documented as of November 2010 (http://projects.tcag.ca/variation/).

On performing the correlation analysis between CNPs and GWAS-SNPs, our results also indicated that several CNPs could be potential causal variants because of their strong correlation with the GWAS-SNPs. Notably, the strong correlation between the CNPs and the GWAS-SNPs near NEGR1 and IRGM for body mass index and Crohn's disease, respectively, are consistent with previous studies.[40,41]

Our study has a higher sensitivity than the study by Díaz de Ståhl *et al.*,[10] which only detected an average of 15 CNVs per individual compared with our study, which detected an average of 36 CNVs per individual. An average of 4 clones per CNV was detected in the Díaz de Ståhl *et al.* study, whereas in our study, each CNV was detected by an average of 51 markers (Supplementary Table 1). The ability to detect smaller CNVs was also demonstrated in our study, because the average size of CNVs detected by Díaz de Ståhl *et al.* was approximately 3.5-fold (358 kb) larger than that in our study. Although Díaz de Ståhl *et al.* also clustered individual overlapping CNVs into loci, their analysis was performed using data from different ancestries (33 Europeans, 24 Africans and 14 Asians), whereas the CNV loci constructed in our study were based entirely on the data from 87 Swedish individuals. Therefore, our list of CNV loci and their frequencies was more representative of the Swedish population.

We did not compare our results with existing data from published studies because of the methodological issues in CNV and ROH detection in the different studies. As different studies have used different platforms, quality control criteria and methods to construct CNV loci and detect ROHs, comparisons with published studies would not be valid. Therefore, we would need to analyze the data from different populations with same analytical procedure. Furthermore, such a comparison is beyond the scope of the current paper and will be addressed in a future publication. However, to provide some preliminary insight into the population differences, we compared the CNP profiles of the Swedish population with the HapMap phase III populations. This comparison was appropriate as

Table 5 Distribution pattern of ROHs across the 22 autosomes

Chromosome	Total number of ROHs	Average number of ROHs per individual	Sum of length of ROHs (bp)	Average sum of length of ROHs per individual (bp)	Chromosome size (bp)[a]	Proportion (%) of chromosome encompassed by ROHs	Number of SNPs in Affymetrix 6.0	Density of SNPs per 10 kb
1	1243	14.3	1 029 256 231	11 830 531	247 249 719	4.78	73469	3.0
2	1491	17.1	1 223 537 523	14 063 650	242 951 149	5.79	75933	3.1
3	1256	14.4	1 069 972 110	12 298 530	199 501 827	6.16	62316	3.1
4	1246	14.3	1 015 875 656	11 676 732	191 273 063	6.10	57561	3.0
5	1021	11.7	859 950 902	9 884 493	180 857 866	5.47	57967	3.2
6	1008	11.6	834 180 388	9 588 280	170 899 992	5.61	57855	3.4
7	811	9.3	632 768 685	7 273 203	158 821 424	4.58	48419	3.0
8	896	10.3	762 529 281	8 764 704	146 274 826	5.99	50019	3.4
9	566	6.5	439 197 494	5 048 247	140 273 252	3.60	42710	3.0
10	722	8.3	612 229 774	7 037 124	135 374 737	5.20	49608	3.7
11	722	8.3	650 352 277	7 475 314	134 452 384	5.56	45944	3.4
12	725	8.3	679 233 723	7 807 284	132 349 534	5.90	43833	3.3
13	482	5.5	360 268 323	4 141 015	114 142 980	3.63	35158	3.1
14	571	6.6	448 210 796	5 151 848	106 368 585	4.84	28942	2.7
15	438	5.0	371 570 656	4 270 927	100 338 915	4.26	26905	2.7
16	192	2.2	159 973 057	1 838 771	88 827 254	2.07	28658	3.2
17	428	4.9	352 288 646	4 049 295	78 774 742	5.14	21347	2.7
18	330	3.8	234 464 335	2 694 992	76 117 153	3.54	27219	3.6
19	184	2.1	143 788 195	1 652 738	63 811 651	2.59	12419	1.9
20	271	3.1	220 116 198	2 530 071	62 435 964	4.05	23487	3.8
21	100	1.1	71 684 424	823 959	46 944 323	1.76	12948	2.8
22	112	1.3	100 622 242	1 156 577	49 691 432	2.33	12059	2.4

Abbreviations: ROHs, regions of homozygosity; SNPs, single-nucleotide polymorphisms; UCSC, University of California Santa Cruz genes.
[a]The size of chromosome was obtained from UCSC Genome Browser.
A full colour version of this Table is available at the Journal of Human Genetics Journal online.

we analyzed the CNP output for the HapMap III populations generated by Canary similar to the Swedish population output. As expected, the results of our analysis showed that the CNP profile of the Swedish population was substantially different from that of the African populations (ASW, MKK and LWK) and CHD. More interestingly, the CNP profile of the Swedish population was also considerably different from that of other European populations (MEX and TSI) and GIH. This further supports the importance of delineating the population characteristics of CNVs/CNPs in the Swedish population.

There are a number of limitations when using SNP genotyping arrays to detect CNVs and ROHs, and the CNV and ROH list reported in our study is not complete. Future studies will require higher sensitivity methods and larger sample sizes for a more thorough detection of CNVs and ROHs. Nevertheless, this is the first population-based study to investigate the population characteristics of CNVs and ROHs in the Swedish population. This study found many reliable CNV loci and also highlighted numerous loci that warrant further investigation for their medical or pharmacogenetic importance. The abundance of ROHs detected in the human genome also suggests the importance of studying their associations with complex phenotypes.

ACKNOWLEDGEMENTS

The Yong Loo Lin School of Medicine, the Life Science Institute and the Office of Deputy President (Research and Technology), National University of Singapore. We also acknowledge the support of the Genome Institute of Singapore, and Agency for Science, Technology and Research, Singapore.

1 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).
2 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
3 Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42,** 400–405 (2010).
4 Yim, S. H., Kim, T. M., Hu, H. J., Kim, J. H., Kim, B. J., Lee, J. Y. *et al.* Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum. Mol. Genet.* **19,** 1001–1008 (2010).
5 Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31,** 851–857 (2010).
6 Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
7 Pinto, D., Marshall, C., Feuk, L. & Scherer, S. W. Copy-number variation in control population cohorts. *Hum. Mol. Genet.* **16,** R168–R173 (2007).
8 Zogopoulos, G., Ha, K. C., Naqib, F., Moore, S., Kim, H., Montpetit, A. *et al.* Germ-line DNA copy number variation frequencies in a large North American population. *Hum. Genet.* **122,** 345–353 (2007).
9 de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P. *et al.* Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* **16,** 2783–2794 (2007).
10 Díaz de Ståhl, T., Sandgren, J., Piotrowski, A., Nord, H., Andersson, R., Menzel, U. *et al.* Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. *Hum. Mutat.* **29,** 398–408 (2008).
11 Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3,** 1787–1799 (2007).
12 Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15,** 789–795 (2006).
13 Li, L. H., Ho, S. F., Chen, C. H., Wei, C. Y., Wong, W. C., Li, L. Y. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* **27,** 1115–1121 (2006).
14 McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Abdel-Rahman, R., Franklin, C. S., Pericic, M. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83,** 359–372 (2008).
15 Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet.* **19,** 2927–2935 (2010).
16 Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl Acad. Sci. USA* **104,** 19942–19947 (2007).
17 Nalls, M. A., Guerreiro, R. J., Simon-Sanchez, J., Bras, J. T., Traynor, B. J., Gibbs, J. R. *et al.* Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10,** 183–190 (2009).
18 Yang, T. L., Guo, Y., Zhang, L. S., Tian, Q., Yan, H., Papasian, C. J. *et al.* Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *J. Clin. Endocrinol. Metab.* **95,** 3777–3782 (2010).

19 O'Dushlaine, C. T., Morris, D., Moskvina, V., Kirov, G., Consortium, I. S., Gill, M. *et al.* Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18,** 1248–1254 (2010).
20 International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
21 Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).
22 Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–1260 (2008).
23 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* **11,** 147 (2010).
24 Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
25 Aqeilan, R. I., Kuroki, T., Pekarsky, Y., Albagha, O., Trapasso, F., Baffa, Re *et al.* Loss of WWOX expression in gastric carcinoma. *Clin. Cancer Res.* **10,** 3053–3058 (2004).
26 Kuroki, T., Yendamuri, S., Trapasso, F., Matsuyama, A., Aqeilan, R. I., Alder, H. *et al.* The tumor suppressor gene WWOX at FRA16D is involved in pancreatic carcinogenesis. *Clin. Cancer Res.* **10,** 2459–2465 (2004).
27 Prickett, T. D., Agrawal, N. S., Wei, X., Yates, K. E., Lin, J. C., Wunderlich, J. R. *et al.* Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. *Nat. Genet.* **41,** 1127–1132 (2009).
28 Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L. *et al*. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* **40,** 1056–1058 (2008).
29 Ouahchi, K., Lindeman, N. & Lee, C. Copy number variants and pharmacogenomics. *Pharmacogenomics* **7,** 25–29 (2006).
30 Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39,** 721–723 (2007).
31 Miki, D., Kubo, M., Takahashi, A., Yoon, K. A., Kim, J., Lee, G. K. *et al.* Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nat. Genet.* **42,** 893–896 (2010).
32 Spencer, K. L., Hauser, M. A., Olson, L. M., Schmidt, S., Scott, W. K., Gallins, P. *et al.* Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. *Hum. Mol. Genet.* **17,** 971–977 (2008).
33 Karypidis, A. H., Olsson, M., Andersson, S. O., Rane, A. & Ekström, L. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *Pharmacogenomics J.* **8,** 147–151 (2008).
34 McCarroll, S. A., Bradner, J. E., Turpeinen, H., Volin, L., Martin, P. J., Chilewski, S. D. *et al.* Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet.* **41,** 1341–1344 (2009).
35 Docampo, E., Rabionet, R., Riveira-Muñoz, E., Escaramís, G., Julià, A., Marsal, S. *et al.* Deletion of the late cornified envelope genes, LCE3C and LCE3B, is associated with rheumatoid arthritis. *Arthritis Rheum.* **62,** 1246–1251 (2010).
36 de Cid, R., Riveira-Munoz, E., Zeeuwen, P. L., Robarge, J., Liao, W., Dannhauser, E. N. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41,** 211–215 (2009).
37 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
38 Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
39 Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
40 Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).
41 McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Journal of Human Genetics (2011) 56, 552–560

www.nature.com/jhg

ORIGINAL ARTICLE

Copy number polymorphisms in new HapMap III and Singapore populations

Chee-Seng Ku[1,2,8], Shu-Mei Teo[1,2,3,8], Nasheen Naidoo[1,2], Xueling Sim[1,2], Yik-Ying Teo[1,2,4,5], Yudi Pawitan[6], Mark Seielstad[7], Kee-Seng Chia[1,2,6] and Agus Salim[1,2,8]

Copy number variations can be identified using newer genotyping arrays with higher single nucleotide polymorphisms (SNPs) density and copy number probes accompanied by newer algorithms. McCarroll *et al*. (2008) applied these to the HapMap II samples and identified 1316 copy number polymorphisms (CNPs). In our study, we applied the same approach to 859 samples from three Singapore populations and seven HapMap III populations. Approximately 50% of the 1291 autosomal CNPs were found to be polymorphic only in populations of non-African ancestry. Pairwise comparisons among the 10 populations showed substantial differences in the CNPs frequencies. Additionally, 698 CNPs showed significant differences with false discovery rate (FDR) < 0.01 among the 10 populations and these loci overlap with known disease-associated or pharmacogenetic-related genes such as *CFHR3* and *CFHR1* (age related macular degeneration), *GSTTI* (metabolism of various carcinogenic compounds and cancers) and *UGT2B17* (prostate cancer and graft-versus-host disease). The correlations between CNPs and genome-wide association studies–SNPs were investigated and several loci, which were previously unreported, that may potentially be implicated in complex diseases and traits were found; for example, childhood acute lymphoblastic leukaemia, age-related macular degeneration, breast cancer, response to antipsychotic treatment, rheumatoid arthritis and type-1 diabetes. Additionally, we also found 5014 novel copy number loci that have not been reported previously by McCarroll *et al*. (2008) in the 10 populations.
Journal of Human Genetics (2011) **56,** 552–560; doi:10.1038/jhg.2011.54; published online 16 June 2011

Keywords: Affymetrix SNP Array 6.0; Birdsuite software; copy number polymorphisms; International HapMap III populations; Southeast Asian populations

INTRODUCTION

The term copy number variation (CNV) was first introduced in 2006 and it is generally defined as additions or deletions in the number of copies of a particular segment of DNA (larger than 1 kb in length) when compared with a reference genome sequence.[1] The ubiquitous nature of CNVs in the human genome was underappreciated until 2004,[2,3] when these reports stimulated a series of efforts to detect and characterise CNVs in different populations.[4–8] This development has also resulted in several new terminologies such as copy number polymorphisms (CNPs), which have been defined as common CNVs with a population frequency of at least 1%.[4]

CNVs can be detected using microarray-based methods, but these have relatively poor resolution when compared with sequencing-based approaches.[9,10] The low resolution of microarray-based methods also led to imprecise mapping of the breakpoints. This is important when constructing copy number loci to estimate population frequencies. Newer genotyping arrays, such as the Illumina Human 1M Beadchip (Illumina, San Diego, CA, USA) and the Affymetrix SNP Arrays 6.0 (Affymetrix, Santa Clara, CA, USA), have higher single nucleotide polymorphisms (SNPs) density and copy number probes, resulting in improved performance of microarray-based methods to detect CNVs. However, even with higher resolution arrays, the challenge of identifying common breakpoints still remains. This is largely due to the early CNV-calling algorithms that identified breakpoints sample-by-sample, resulting in significant variation of breakpoints. The Canary algorithm in the Birdsuite software overcomes this problem by calling CNPs simultaneously across multiple individuals at pre-defined genomic locations.[11] McCarroll *et al*.[4] used the Canary algorithm to identify 1316 CNPs in the HapMap Phase II populations. These CNPs were well validated and their sizes were in agreement with the results from the fosmid paired-end sequencing experiment.[9]

[1]Centre for Molecular Epidemiology, National University of Singapore, Singapore, Singapore; [2]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore, Singapore; [4]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore; [5]Department of Statistics & Applied Probability, National University of Singapore, Singapore, Singapore; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and [7]Laboratory Medicine, Institute of Human Genetics, University of California, San Francisco, CA, USA
Correspondence: C-S Ku or Assistant Professor A Salim, Centre for Molecular Epidemiology, Department of Epidemiology and Public Health (MD3), Yong Loo Lin School of Medicine, National University of Singapore, 16 Medical Drive, Singapore 117597, Singapore.
E-mail: g0700040@nus.edu.sg or ephaguss@nus.edu.sg
[8]These authors contributed equally to this work.
Received 26 November 2010; revised 3 May 2011; accepted 6 May 2011; published online 16 June 2011

To provide a more global map of CNPs, our study aims to determine integer copy numbers of the 1316 CNPs set of three Southeast Asian populations in Singapore, namely Chinese (Sing–Chinese), Malay (Sing–Malay) and Asian Indian (Sing–Indian), and the seven populations from the HapMap Phase III.[12] The HapMap III populations studied are people of African ancestry in the southwestern USA (ASW), the Chinese community in Metropolitan Denver, Colorado, USA (CHD), Gujarati Indians in Houston, Texas, USA (GIH), the Luhya in Webuye, Kenya (LWK), people of Mexican ancestry in Los Angeles, California, USA (MEX), the Maasai in Kinyawa, Kenya (MKK) and the Tuscans in Italy (TSI). The characteristics of CNPs in the 10 populations will be described and compared. In addition, the correlation between CNPs and SNPs in the 10 populations will also be characterised and compared. A special emphasis will be given to studying the correlation between SNPs in the genome-wide association studies (GWAS) catalog (GWAS–SNPs) and CNPs in the 10 populations. Additionally, novel copy number loci that have not been reported previously by McCarroll *et al*.[4] will also be reported on from the 10 populations.

MATERIALS AND METHODS

DNA samples and genotyping

The detailed information on the sources of DNA samples, demographic data of the samples, sample selection and the origin and migration history of the three Singapore populations (Chinese, Malay and Asian Indian) have been described in our previous publication.[8,13] This study was approved by the National University of Singapore Institutional Review Board (Reference Code: 07-199E). In total, 292 DNA samples (99 Chinese, 98 Malay and 95 Indian) were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Of the 292 samples, 27 were excluded from subsequent analysis. The final set of 265 samples (93 Chinese, 88 Malays and 84 Indians) was available for analysis using Birdsuite. There were 135 females and 130 males in the final dataset. The detailed information on the quality control and sample filtering have also been described in our previous papers.[8,13]

HapMap III samples

The CEL-files of the Affymetrix SNP Array 6.0 for the seven populations in HapMap III were downloaded from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). All the samples were analysed by Birdsuite, with only unrelated samples included in our study; that is, family-related samples were removed using the 'relationships' file provided by the International HapMap Project. After the sample exclusion step, a total of 594 unrelated samples from the seven HapMap III populations were analysed: ASW (n=52), CHD (n=89), GIH (n=89), LWK (n=90), MEX (n=53), MKK (n=132) and TSI (n=89).

CNP calling using Canary

The Birdsuite software was used to analyse the Affymetrix SNP Array 6.0 dataset, which consisted of two components for detecting copy number changes. The first component, Canary, was used to determine the integer copy number at each of the predefined 1316 CNPs identified by McCarroll *et al*.[4] in the HapMap II samples. These CNPs were found in more than one HapMap II individual and the sizes of these CNPs were also determined. The 1316 CNPs were distributed in all the autosomes and sex chromosomes. However, 25 CNPs located in the sex chromosomes were removed, as CNP calling in sex chromosomes is more problematic and less accurate. Therefore, the results reported in this study comprised of only 1291 CNPs in the 22 autosomes. Confidence statistics was used to identify poor quality calls and only integer copy numbers detected with high confidence (confidence score <0.1) were reported and used for subsequent analyses. We performed the Hardy–Weinberg equilibrium analysis as a quality control measure for biallelic CNPs in all 10 populations. It is recommended that the samples should be analysed on the basis of the genotyping batches using Birdsuite; therefore, the samples for Singapore and HapMap III populations were analysed by batch without separating the samples into each specific population.

FDR correction for population comparisons of the integer copy numbers of the CNPs

Population differences in the integer copy numbers were examined using the Fisher's exact test as implemented by the 'fisher test' command in R. The false discovery rate (FDR) was used in place of the P-value to account for the multiple-testing problem. We calculated the FDR using the Benjamini and Hochberg method. We performed two different test procedures: (1) comparing the integer copy numbers among the 10 populations simultaneously and (2) pairwise comparisons of the integer copy numbers among the 10 populations. For each procedure, FDR was computed once to control for all the tests (that is, in the second procedure, we calculated the FDR once by combining the P-values from 45×1291 tests).

Correlation analysis

All the correlation analyses of CNPs and nearby SNPs were done separately for each of the 10 populations. For each autosomal CNP (restricted to biallelic CNPs with MAF$\geqslant$5%), SNPs in close proximity with the CNP; that is, within a 200-kb window from the start- and end-position of the CNP were considered. The square of the Pearson correlation coefficient (r^2) for each of the SNPs (excluding the SNPs used for CNP-calling) found within the 200-kb windows of the respective CNP was then calculated.

The r^2 is the square of the Pearson correlation coefficient between the copy number genotypes and the SNP genotypes. The copy number genotypes were obtained using Canary in the Birdsuite algorithm. The SNP genotypes were obtained using Larry Bird in the Birdsuite algorithms. Larry Bird outputs the number of allele A (0, 1, 2) and number of allele B (0, 1, 2) for each SNP. We used the number of allele A for the calculation. Larry Bird generates the number of allele A and number of allele B for each SNP. As each SNP has two alleles in total, knowing the number of allele A will inform the number of allele B; for example, if the number of allele A is 2, then number of allele B should be 0.

The same r^2 calculations used for the autosomal CNPs and the SNPs identified by GWAS were used to explore the potential associations of CNPs with human diseases and traits. The list of GWAS–SNPs was downloaded from the National Human Genome Research Institute's website (http://www.genome.gov/gwastudies/) on 24 May 2010.

Copy number loci calling using Birdseye and validation

The Birdseye component in Birdsuite was used to detect additional copy number loci located outside the 1316 CNPs in the 10 populations. Similarly, only the copy number loci in autosomal chromosomes were detected because of the inaccuracy of Birdseye in detecting copy number loci in the sex chromosomes. Copy number calls with low confidence (confidence score <5) were removed. On the basis of the copy number calls generated by Birdseye, we constructed novel copy number loci using the methods that we developed previously.[14] All the downstream analyses after Canary and Birdseye were performed using the software package R (http://www.r-project.org/). The novel copy number loci identified by Birdseye were compared with data from the Database of Genomic Variants (http://projects.tcag.ca/variation/) as a validation step. We defined a copy number locus overlapped with the Database of Genomic Variants, if the locus overlapped by $>50\%$ of its length with one or more entries in the Database of Genomic Variants.

RESULTS

Characteristics of CNPs in the 10 populations

In each of the 10 populations, among the polymorphic CNPs (Table 1), most were biallelic, where the integer copy numbers were either exclusively deletions (copy number=0, 1) or exclusively duplications (copy number=3, 4). Among the biallelic CNPs, the majority did not show significant deviation from Hardy–Weinberg equilibrium with less than 2% failing a Hardy–Weinberg equilibrium test at P-value <0.01 in all except three populations—Sing–Chinese (2.2%), ASW (4.2%) and LWK (2.8%).

Table 1 The number of loci (and the percentage) with varying population frequencies for the 1291 autosomal CNPs

CNP	*Sing–Chinese*	*Sing–Malay*	*Sing–Indian*	*ASW*	*CHD*	*GIH*	*LWK*	*MEX*	*MKK*	*TSI*
Not polymorphic (0%)	675[a] (52.29)[b]	663 (51.36)	670 (51.90)	341 (26.41)	688 (53.33)	677 (52.44)	487 (37.72)	681 (52.75)	460 (35.63)	650 (50.35)
Population frequencies $\leqslant$10%	335 (25.95)	342 (26.49)	324 (25.10)	592 (45.86)	330 (25.58)	318 (24.63)	458 (35.48)	336 (26.03)	507 (39.27)	355 (27.50)
Population frequencies $>$10–50%	155 (12.01)	158 (12.24)	170 (13.17)	242 (18.75)	141 (10.93)	174 (13.48)	229 (17.74)	152 (11.77)	208 (16.11)	157 (12.16)
Population frequencies $>$50%, $<$100%	109 (8.44)	113 (8.75)	109 (8.44)	103 (7.98)	109 (8.45)	106 (8.21)	101 (7.82)	105 (8.13)	99 (7.67)	113 (8.75)
Completely polymorphic (100%)	17 (1.32)	15 (1.16)	18 (1.39)	13 (1.01)	22 (1.71)	16 (1.24)	16 (1.24)	17 (1.32)	17 (1.32)	16 (1.24)

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.
[a]Number of loci.
[b]Percentage (number of loci/1291 autosomal CNPs).

In terms of the proportion of non-polymorphic loci and loci with varying population frequencies, the Singapore populations were similar to the HapMap III populations of non-African descent (CHD, GIH, MEX and TSI) (Table 1 and Supplementary Figure 1). More than half of the CNPs were non-polymorphic in the Singapore and HapMap III populations of non-African descent. This was in contrast to the populations of African descent (ASW, LWK and MKK), where only 26.41–37.72% of the CNPs were not polymorphic. They also had higher proportions of CNPs with frequencies ranging from 1 to 10%, ASW (45.86%), LWK (35.48%) and MKK (39.27%), compared with the other populations (ranging from 24.63 to 27.50%). In addition, among all the populations, there were no substantial differences in the proportion of CNPs with a population frequency $>$10%. The discrepancy between populations of African descent and others is largely due to these populations having a larger number of rarer CNPs with a population frequency $<$10%. Hence, the differences between populations of African descent and the others were primarily in the proportion of non-polymorphic loci and those with population frequencies $<$10%. It is also worth noting that the Sing–Indian and Sing–Chinese populations have almost similar distributions of polymorphic loci, when compared with the HapMap III populations with whom they share a similar ancestry (that is, GIH and CHD, respectively) (Table 1 and Supplementary Figure 1).

The proportion of common (MAF$\geqslant$0.05) biallelic CNPs that were highly correlated with at least one SNP ($r^2>0.8$) was approximately 50% for non-African populations, but a lower proportion for African populations; that is, ASW (35.34%), LWK (34.84%) and MKK (37.39%). The majority of the common biallelic CNPs were 'deletions'. There was a substantial difference in the proportion that was highly correlated with at least one SNP for CNPs categorised as 'deletions' and 'duplications'. However, this substantial difference could be biased because of the small number of 'duplications' (Table 2). The strength of correlation or the r^2 value decreased with distance between the CNP and SNP (Supplementary Figure 2).

We further investigated whether CNPs that were not well tagged were located in the genomic regions where SNP markers are sparse. The correlation patterns did not seem to be affected by the number of nearby SNPs and the MAF of CNPs. There was no apparent difference in the number of nearby SNPs and the MAF of CNPs between (a) the CNPs that were in strong correlation ($r^2>0.8$) and (b) CNPs that were not in strong correlation with SNPs (Supplementary Figures 3a and b). However, smaller sizes of CNPs were generally in strong correlation with more SNPs than the larger CNPs (Supplementary Figure 3c). These results were consistent across the 10 populations.

Population differences in the integer copy numbers of the CNPs

Out of the 698 CNPs (FDR$<$0.01) that differed between the 10 populations, several loci encompassed known disease- or traits-associated or pharmacogenetic-related genes (Supplementary Table 1). These included *WWOX*, *ERBB4* and *TP63* (cancers), *ADAMTSL3* (height), *CFHR3* and *CFHR1* (age-related macular degeneration), *GSTT1* (metabolism of various carcinogenic compounds and cancers), *UGT2B17* (prostate cancer and graft-versus-host disease) and *CYP2A6* (metabolism of various drugs). There was a large interpopulation difference in the frequencies of some of the CNPs overlapping these genes. For example, CNP2203, which overlaps with the tumour suppressor gene *WWOX*, was not polymorphic in CHD, whereas it had a deletion frequency of 2.38% in Sing–Chinese and 7.32% in Sing–Malay (Table 3 and Supplementary Table 2). In contrast, the deletion frequency was 51.81% in Sing–Indian and 48.86% in GIH. Similarly, CNP147, which overlaps with the *CFHR3* and *CFHR1* genes, had

Table 2 The number and proportion (%) of common (MAF ⩾ 0.05) biallelic (a) CNPs, (b) deletions, (c) duplications that were highly correlated with at least one SNPs ($r^2 > 0.8$)

Population	No. of CNPs (MAF⩾5%)	No. of CNPs correlated ($r^2 > 0.8$)	Proportion (%)	No. of deletions (MAF⩾5%)	No. of deletions correlated ($r^2 > 0.8$)	Proportion (%)	No. of duplications (MAF⩾5%)	No. of duplications correlated ($r^2 > 0.8$)	Proportion (%)
Sing–Chinese	194	104	53.61	174	103	59.20	20	1	5.00
Sing–Malay	190	106	55.79	170	105	61.76	20	1	5.00
Sing–Indian	210	115	54.76	190	112	58.95	20	3	15.00
ASW	266	94	35.34	241	94	39.00	25	0	0.00
CHD	201	112	55.72	181	110	60.77	20	2	10.00
GIH	216	117	54.17	197	117	59.39	19	0	0.00
LWK	263	89	33.84	242	87	35.95	21	2	9.52
MEX	229	105	45.85	204	104	50.98	24	1	4.17
MKK	230	86	37.39	210	86	40.95	20	0	0.00
TSI	205	105	51.22	183	103	56.28	22	2	9.09

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MAF, minor allele frequency; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; SNPs, single-nucleotide polymorphisms; TSI, Tuscans in Italy.
r^2, Square of the Pearson correlation coefficient.

Table 3 CNPs (FDR < 0.01) that overlap with known disease-associated or pharmacogenetic-related genes

CNP	Gene	Sing–Chinese	Sing–Malay	Sing–Indian	ASW	CHD	GIH	LWK	MEX	MKK	TSI
CNP2203	*WWOX*	2.38[a]	7.32	51.81	66.67	0.00	48.86	40.00	67.31	28.35	68.18
CNP340	*ERBB4*	0.00	2.33	12.05	7.69	0.00	17.24	0.00	0.00	0.00	4.49
CNP530	*TP63*	64.84	48.24	27.38	30.77	68.54	31.82	31.82	9.62	32.06	6.90
CNP2118	*ADAMTSL3*	67.05	46.84	11.54	38.46	51.19	4.49	49.40	24.32	48.80	19.51
CNP147	*CFHR3, CFHR1*	11.83	12.64	53.57	59.62	15.73	58.43	59.09	18.87	42.42	43.82
CNP2560	*GSTT1*	96.77	85.06	56.63	72.00	92.13	70.79	75.56	71.70	80.15	67.06
CNP603	*UGT2B17*	100.00	95.40	82.14	48.08	98.88	86.42	63.33	58.49	67.18	58.43
CNP2415	*CYP2A6*	18.89	36.25	5.13	6.00	23.86	11.49	8.05	2.04	8.80	4.60

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; FDR, false discovery rate; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.
[a]Population frequency (%)=deletion frequency+duplication frequency.

deletion frequencies in Sing–Chinese (10.75%), Sing–Malay (12.64%) and CHD (15.73%) that was substantially lower than the other populations.

Another CNP of interest was CNP2560, a 46-kb deletion that overlaps with *GSTT1*. *GSTT1* is an important detoxification enzyme and has a key role in metabolism of carcinogenic compounds. The total deletion frequency of this CNP was high in all the 10 populations ranging from 56.63 to 96.77% (Table 3 and Supplementary Table 2). Sing–Indians had a considerably lower total deletion frequency (56.63%) than Sing–Malays (85.06%) and Sing–Chinese (96.77%). This difference is attributable to two-copy deletion, as the difference in two-copy deletion frequency ranged from 15.66% in Sing–Indian, 32.18% in Sing–Malay and 46.24% in Sing–Chinese. The two Chinese populations had the highest two-copy deletion frequency (CHD, 41.57%). Conversely, both the Indian populations had the lowest two-copy deletion frequency (GIH, 17.98%).

CNP603 is a 125-kb deletion that overlaps with *TMPRSS11E* and *UGT2B17*. The entire *UGT2B17* gene is within the deletion locus, but only one exon from *TMPRSS11E* was deleted. The deletion frequency of CNP603 was very different in Asian and non-Asian populations (Table 3 and Supplementary Table 2). Asian populations (Sing–Chinese, Sing–Malay, Sing–Indian, CHD and GIH) had higher frequencies, which ranged from 82.14 to 100%, when compared with populations of European and African ancestry (48.08–67.18%). The differences were even more apparent for two-copy deletions with the highest frequencies in CHD (70.79%), Sing–Chinese (65.59%) and Sing–Malay (52.87%), followed by the two Indian populations, GIH (37.04%) and Sing–Indian (30.95%), whereas the European and African populations were in the lower end of the spectrum with frequencies <20%. Generally, this trend was reversed for the frequency of one-copy deletions especially in the Singapore populations (Sing–Chinese 33.33%, Sing–Malay 42.53% and Sing–Indian 51.19%).

The number of CNPs that showed significant differences (FDR < 0.01) in pairwise comparisons of the 10 populations are shown in Table 4. Only 19 CNPs showed significant differences between Sing–Chinese and CHD, and 12 CNPs between Sing–Indian and GIH, suggesting that the CNPs profile in the two Chinese and two Indian populations were very similar (Supplementary Figure 4). Through these pairwise comparisons (Table 4 and Supplementary Figure 4), the 10 populations can be divided into three groups representing Asian, European and African ancestry: (a) Sing–Chinese, Sing–Malay and CHD, (b) Sing–Indian, GIH, MEX and TSI, (c) ASW, LWK and MKK. The CNPs profiles of Sing–Indian and GIH were closer to European populations (MEX and TSI).

Correlation analysis between CNPs and GWAS-SNPs

To investigate the potential role of CNPs in the aetiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs in

Table 4 The number of CNPs that showed significant differences (FDR < 0.01) in the pairwise comparisons among the 10 populations

Population	*Sing–Chinese*	*Sing–Malay*	*Sing–Indian*	*ASW*	*CHD*	*GIH*	*LWK*	*MEX*	*MKK*	*TSI*
Sing–Chinese	—	6	84	137	19	106	209	81	199	141
Sing–Malay	—	—	46	125	26	72	197	59	180	126
Sing–Indian	—	—	—	93	88	12	186	32	147	54
ASW	—	—	—	—	132	95	13	69	18	90
CHD	—	—	—	—	—	113	196	77	192	130
GIH	—	—	—	—	—	—	170	35	155	52
LWK	—	—	—	—	—	—	—	123	33	176
MEX	—	—	—	—	—	—	—	—	97	27
MKK	—	—	—	—	—	—	—	—	—	146
TSI	—	—	—	—	—	—	—	—	—	—

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; FDR, false discovery rate; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.

the National Human Genome Research Institute GWAS catalog. Out of the >2500 GWAS–SNPs that have been found to be associated with various complex diseases and traits, only 17 GWAS–SNPs were found to be in strong correlation with 12 CNPs (Table 5 and Supplementary Tables 3 and 4). In this analysis, we defined a strong correlation as $r^2 > 0.5$, following Conrad *et al*.[5] These 17 SNPs were reported to be associated with 14 diseases or traits and the notable phenotypes that were observed consistently across the populations were body mass index, Crohn's disease, multiple sclerosis, myocardial infarction and prostate cancer. Several SNPs were in strong correlation with a single CNP; for example, three SNPs (rs13361189, rs1000113, rs11747270) were found to be in strong correlation with CNP874. Of the 33 copy number loci identified by Conrad *et al*.,[5] which were in strong correlation with GWAS–SNPs, seven were also identified in our study which had >50% overlap in length. The remaining five CNPs in our study were associated with childhood acute lymphoblastic leukaemia, age-related macular degeneration, breast cancer, response to antipsychotic treatment, rheumatoid arthritis and type-1 diabetes (Table 5 and Supplementary Tables 3 and 4).

Several SNPs were consistently found to be in strong correlation with four CNPs (CNP60, CNP874, CNP877 and CNP333) in all populations. The most notable was rs2815752 near the *NEGR1* gene (associated with body mass index), which is in perfect correlation ($r^2=1$) with CNP60 in all the 10 populations (Table 5 and Supplementary Table 3). This locus is a 42-kb deletion located in chromosome 1, which did not overlap with any of the UCSC (University of California, Santa Cruz) genes and it is located only 1.3 kb away from the SNP. The total deletion frequency in the three Singapore populations was high (Figure 1a and Supplementary Table 5). There were, however, differences in the frequency of two-copy deletion. More than 80% of the Sing–Chinese and Sing–Malay samples were deleted in both copies, but only about 41% for the Sing–Indian samples. The pattern is similar between Sing–Chinese and CHD, as well as Sing–Indian and GIH. The frequency of two-copy deletion frequency varied substantially across the 10 populations, from the lowest in the LWK population (26.97%) to the highest in Sing–Chinese (87.10%). A significant difference in the two-copy deletion frequency of CNP60 was seen between Asian populations (>80% for Sing–Chinese, Sing–Malay and CHD) compared with African populations (<35% for ASW, LWK and MKK), whereas the frequency of the Sing–Indian and GIH resembles European populations (MEX and TSI) (Supplementary Table 5).

CNP874 was found to be in strong correlation with three GWAS–SNPs located near the *IRGM* gene, which is associated with Crohn's disease. This strong correlation pattern was consistent across the 10 populations (Table 5). Most of the individuals carried either deletions or had a diploid copy. This locus spans 13 kb in chromosome 5 and did not overlap with any of the UCSC genes. The SNPs were located 4.8 kb (rs13361189), 21.4 kb (rs1000113) and 40.2 kb (rs11747270) away from the deletion. The differences in the frequency of two-copy deletion of CNP874 appeared to divide the 10 populations into two clusters. The populations of European ancestry (MEX and TSI) and Indian populations (Sing–Indian and GIH) had a frequency ⩽6.41%, but the other populations had higher frequencies, which ranged from 10% to 20.69% (Figure 1b and Supplementary Table 5). We also found a substantially lower frequency of two-copy deletion in the Sing–Indian (6.41%) compared with the Sing–Chinese (15.22%) and the Sing–Malay (11.49%) populations.

The CNP877 locus has been implicated in multiple sclerosis. It was however not polymorphic in the Sing–Chinese (Figure 1c and Supplementary Table 5). The total deletion frequencies for Sing–Malay and CHD were 2.30 and 1.14%, respectively. However, we found a much higher total deletion frequency for the other seven populations, which ranged from 17.05 to 42.53%.

Novel copy number loci in the 10 populations

The second component of the Birdsuite software, Birdseye, was used to identify novel copy number loci in the 10 populations. We subsequently found 5947 copy number loci, of which 933 loci were excluded because of overlap with the 1291 autosomal CNPs identified by McCarroll *et al*.[4] As a result, only 5014 were novel copy number loci; that is, had not been previously found by McCarroll *et al*.[4] Of these, 1448 loci were detected in two or more individuals in the 10 populations (Table 6). The list of these loci is available in Supplementary Table 6. Using a more stringent definition of 'common' novel copy number loci (population frequency ⩾1%), there were only 170 loci and of these, 42 loci had a population frequency ⩾5%.

Of the 1448 novel copy number loci, 763 (52.69%) were found to overlap with the data from the Database of Genomic Variants. Although for the 170 loci, the overlap was 78.82% (Table 6). Additionally, we also found that 86.54% of the 1448 loci were biallelic; that is, these loci contained either deletions (48.76%) or duplications (37.78%). The remaining loci were found to have both deletions and duplications. The majority of these loci did not overlap with the UCSC genes (62.43%). Of the 170 loci, 37.06% contained both deletions and duplications and the majority of these loci also did not overlap with the UCSC genes (52.35%).

Table 5 Correlation between CNPs and GWAS–SNPs at $r^2 > 0.5$ in 10 populations

CNP	*Chr.*	*Start/end position*	*GWAS–SNPs*	*GWAS–SNPs position*	*Population*	*Gene*	*Disease/trait*
60	1	72 541 504 72 583 736	rs2815752	72 585 028	Sing–Chinese, Sing–Malay, Sing–Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	*NEGR1*	BMI
874	5	150 185 693 150 198 797	rs13361189	150 203 580	Sing–Chinese, Sing–Malay, Sing–Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	*IRGM*	Crohn's disease
874	5	150 185 693 150 198 797	rs1000113	150 220 269	Sing–Chinese, Sing–Malay, Sing–Indian, CHD, MEX, MKK, TSI	*IRGM*	Crohn's disease
874	5	150 185 693 150 198 797	rs11747270	150 239 060	Sing–Chinese, Sing–Malay, Sing–Indian, ASW, CHD, GIH, MEX, MKK, TSI	*IRGM*	Crohn's disease
877	5	155 409 350 155 415 307	rs4704970	155 433 570	Sing–Malay, Sing–Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	*SGCD*	Multiple sclerosis
333	2	203 608 045 203 610 291	rs6725887	203 454 130	Sing–Chinese, CHD, LWK, MEX, MKK, TSI	*WDR12*	Myocardial infarction (early onset)
399	3	37 957 108 37 961 932	rs9311171	37 971 481	Sing–Chinese, Sing–Malay, CHD, MEX, TSI	*CTDSPL*	Prostate cancer
28	1	25 465 715 25 534 592	rs10903129	25 641 524	Sing–Indian, GIH	*TMEM57*	Total cholesterol
147	1	194 997 658 195 068 695	rs6428370	195 111 216	Sing–Indian, ASW, GIH, MEX, TSI	*Intergenic*	Acute lymphoblastic leukaemia (childhood)
147	1	194 997 658 195 068 695	rs10737680	194 946 078	GIH	*CFH*	Age-related macular degeneration
1491	9	98 700 200 98 729 161	rs10816533	98 578 959	CHD	*ZNP510*	Height
109	1	150 822 330 150 853 218	rs10888501	150 804 578	Sing–Malay, Sing–Indian	*Intergenic*	Response to antipsychotic treatment
12035	12	118 473 270 118 475 144	rs11064768	118 302 892	Sing–Chinese	*CCDC60*	Schizophrenia
2197	16	72 953 795 73 009 537	rs10871290	73 030 197	Sing–Indian	*GLG1*	Breast cancer
933	6	32 539 530 32 681 749	rs3135338	32 509 195	Sing–Malay, Sing–Indian	*HLA*	Multiple sclerosis
933	6	32 539 530 32 681 749	rs615672	32 682 149	Sing–Malay	*HLA-DRB1*	Rheumatoid arthritis
933	6	32 539 530 32 681 749	rs9272346	32 712 350	Sing–Malay	*MHC*	Type 1 diabetes

Abbreviations: ASW, African ancestry in the southwestern USA; BMI, body mass index; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; GIH, Gujarati Indians in Houston, Texas, USA; GWAS–SNP, genome-wide association studies–single nucleotide polymorphisms; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.
r^2, Square of the Pearson correlation coefficient.

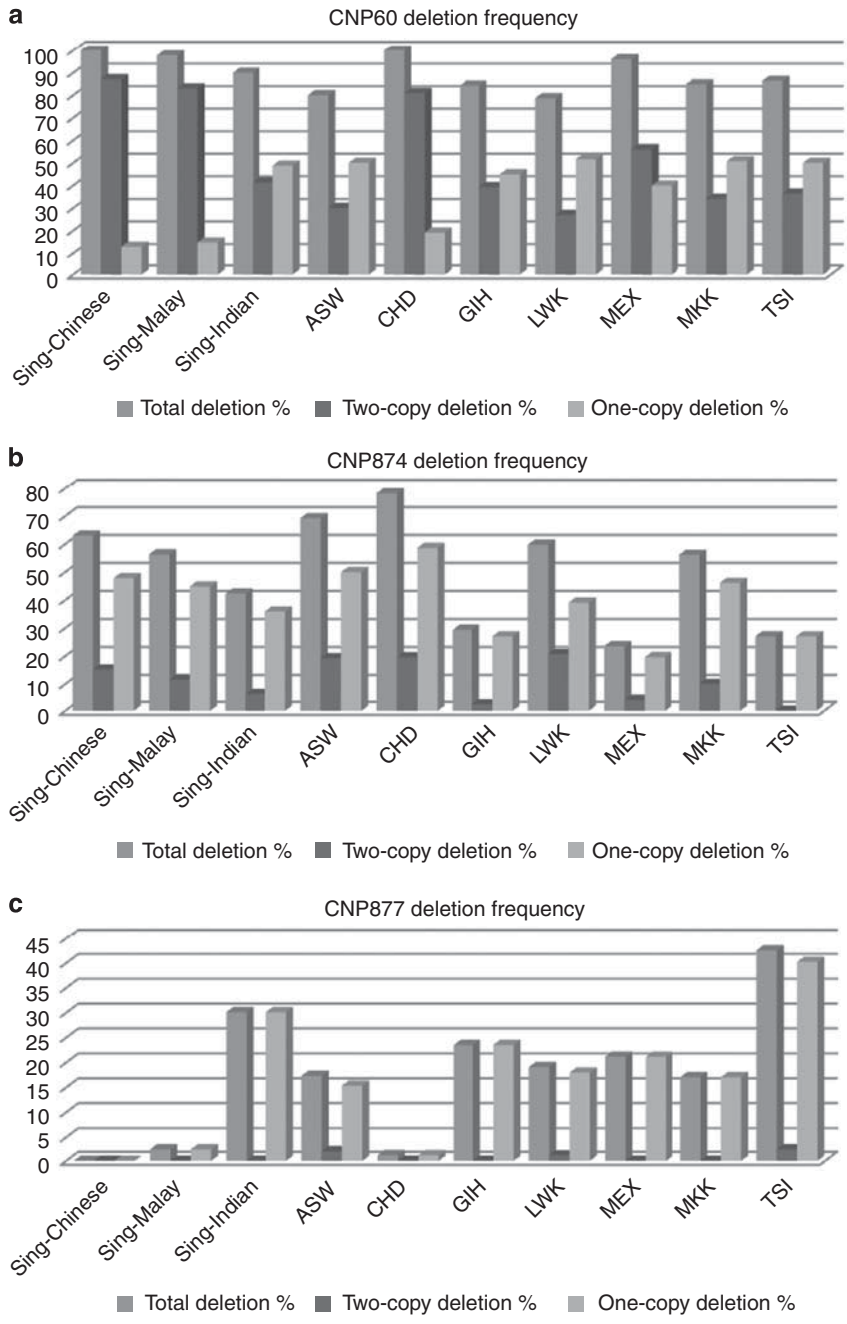


Figure 1 Total, two-copy and one-copy deletion frequencies of (**a**) CNP60, (**b**) CNP874 and (**c**) CNP877 in 10 populations.

DISCUSSION

The finding that approximately 50% of the CNPs identified by the McCarroll *et al*.[4] study were not polymorphic in all of the three Singapore populations and the HapMap III populations (CHD, GIH, MEX and TSI) suggests that the CNPs found in the 'reference' HapMap II populations are not necessarily polymorphic or common in other populations. This finding, together with the identification of novel copy number loci other than those found using the HapMap II populations, highlights the importance of characterising CNPs in different populations.

In addition, we also found several hundred CNPs that showed significant differences in integer copy numbers among the 10 populations. More interestingly, many of these loci encompass genes of medical relevance. For example, we found a markedly lower deletion frequency at CNP2203 (which is associated with the *WWOX* gene) in Sing–Chinese and Sing–Malay compared with other populations. *WWOX* is a tumour suppressor gene affected in multiple cancers.[15] On the other hand, deletion of the *UGT2B17* gene was also been found to be associated with an increased risk of prostate cancer.[16,17] The functional role of the *UGT2B17* enzyme is clear in prostate cancer, as it is involved in steroid hormone (androgen) metabolism. The mismatch of *UGT2B17* copy numbers in donors and recipients of stem cell transplantation were also associated with an increased risk of graft-versus-host disease.[18] This gene is contained within CNP603, which

Table 6 Characteristics of novel copy number loci identified in 10 populations using Birdseye

Detail	*Number (%)*	
General characteristics		
Novel copy number loci constructed from Birdseye	5014	
Number of loci that detected in ⩾two individuals	1448 (28.88)	
Number of loci that detected in ⩾1% of the studied sample size (859 samples); that is, detected in ⩾eight individuals	170 (3.39)	
Number of loci ⩾5%; that is, detected in ⩾43 individuals	42 (0.84)	
Focus on the loci detected (A) in ⩾two individuals and (B) in ⩾1% of the studied sample size	*(A) (n=1448 loci)*	*(B) (n=170 loci)*
Sum of the total length (Mb)	232.78	65.98
Average length per locus (kb)	160.76	388.11
Average number of markers per locus	82	143
Size distribution of loci		
<1 kb	56 (3.87)	5 (2.94)
1–<10 kb	325 (22.44)	31 (18.24)
10–<50 kb	420 (29.01)	46 (27.06)
50–<100 kb	165 (11.40)	10 (5.88)
100–<500 kb	354 (24.45)	35 (20.59)
500 kb–<1 Mb	91 (6.28)	22 (12.94)
>1 Mb	37 (2.56)	21 (12.35)
Deletion or duplication status		
Loci with only deletion	706 (48.76)	78 (45.88)
Loci with only duplication	547 (37.78)	29 (17.06)
Loci with deletion and duplication	195 (13.47)	63 (37.06)
Overlapping with DGV		
Loci that overlap with ⩾50% with the DGV	763 (52.69)	134 (78.82)
Loci that did not overlap with DGV	685 (47.31)	36 (21.18)
Overlapping with UCSC genes		
Loci that overlap with UCSC genes	544 (37.57)	81 (47.65)
Loci that did not overlap with UCSC genes	904 (62.43)	89 (52.35)

Abbreviations: DGV, Database of Genomic Variants; UCSC, University of California, Santa Cruz.

show substantial differences between the Singapore and HapMap III populations. Although a direct association between the CNPs and phenotypic differences is not established in our study, collectively our results suggest that CNPs distributions are substantially different between populations and thus, may account for phenotypic differences between them.

We found 12 CNPs that may have potential implications in various diseases and traits; however, only five of them have not been reported by Conrad *et al.*,[5] who found evidence of correlations for 33 copy number loci with GWAS–SNPs at $r^2>0.5$. The difference in the number of loci found to be in correlation with GWAS–SNPs between our study and the Conrad *et al.*[5] study is likely due to the limitation that we only focused on the 1291 CNPs, whereas Conrad *et al.*[5] studied the whole genome. Furthermore, it could also be due to the difference in the marker density of the microarrays used in our study and the Conrad *et al.*[5] study. We used the Affymetrix SNP Array 6.0, whereas they used a set of 20 oligonucleotide-CGH arrays, comprising 42 million probes. The differences in marker density will contribute to the differences in sensitivity of detection.[5]

Several previous studies have reported correlations between CNVs and GWAS–SNPs. For example, deletions near *IRGM* and *NEGR1* genes, which were in perfect linkage disequilibrium (LD) with the GWAS–SNPs, were identified for Crohn's disease and body mass index, respectively.[19,20] Our study also showed strong correlations between CNPs and GWAS–SNPs near *IRGM* and *NEGR1* in all 10 populations, but the deletion frequencies varied substantially among the populations. GWAS–SNPs are potentially indirect markers of disease variants, which include CNPs. This may have important clinical implications if these deletions are true disease variants.

A recent paper published by the International HapMap Consortium also studied CNPs in the HapMap III populations.[12] However, they merged and analysed the probe-level intensity data from both the Affymetrix SNP Array 6.0 and the Illumina 1M Beadchip arrays. In contrast, we only analysed the Affymetrix SNP Array 6.0 data and focused primarily on the 1291 CNPs identified previously, as only the raw signal intensity files of this array were available from the HapMap website. A total of 1610 CNPs with an estimated frequency of at least 1% of the cohort were identified in the HapMap III populations by the International HapMap Consortium. They also found that most CNPs also occurred at a low frequency.[12] This was consistent with our study where among the polymorphic CNPs, the majority also occurred at a low frequency (<10%). Similarly, the finding that the frequency spectrum of common CNPs (>10%) was similar across populations by the International HapMap Consortium was in good agreement with our results (Table 1).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was supported by the Yong Loo Lin School of Medicine, the Life Science Institute and the Office of Deputy President (Research and Technology), National University of Singapore. We also acknowledge the technical and financial support of the Genome Institute of Singapore and Agency for Science, Technology and Research, Singapore.

1 Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16,** 949–961 (2006).
2 Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
3 Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
4 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).
5 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
6 Yim, S. H., Kim, T. M., Hu, H. J., Kim, J. H., Kim, B. J., Lee, J. Y. *et al.* Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum. Mol. Genet.* **19,** 1001–1008 (2010).
7 Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42,** 400–405 (2010).
8 Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31,** 851–857 (2010).
9 Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).

10 Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
11 Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–1260 (2008).
12 International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
13 Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19,** 2154–2162 (2009).
14 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of Copy-Number Variants across multiple individuals. *BMC Bioinformatics* **11,** 147 (2010).
15 Lewandowska, U., Zelazowski, M., Seta, K., Byczewska, M., Pluciennik, E. & Bednarek, A. K. WWOX, the tumour suppressor gene affected in multiple cancers. *J. Physiol. Pharmacol.* **60,** 47–56 (2009).
16 Park, J., Chen, L., Ratnashinge, L., Sellers, T. A., Tanner, J. P., Lee, J. H. *et al.* Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. *Cancer Epidemiol. Biomarkers Prev.* **15,** 1473–1478 (2006).
17 Karypidis, A. H., Olsson, M., Andersson, S. O., Rane, A. & Ekström, L. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *Pharmacogenomics J.* **8,** 147–151 (2008).
18 McCarroll, S. A., Bradner, J. E., Turpeinen, H., Volin, L., Martin, P. J., Chilewski, S. D. *et al.* Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet.* **41,** 1341–1344 (2009).
19 McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).
20 Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Hum Genet (2011) 129:1–15
DOI 10.1007/s00439-010-0920-6

REVIEW PAPER

Regions of homozygosity and their impact on complex diseases and traits

Chee Seng Ku · Nasheen Naidoo · Shu Mei Teo · Yudi Pawitan

Received: 8 August 2010 / Accepted: 4 November 2010 / Published online: 23 November 2010

Abstract Regions of homozygosity (ROHs) are more abundant in the human genome than previously thought. These regions are without heterozygosity, i.e. all the genetic variations within the regions have two identical alleles. At present there are no standardized criteria for defining the ROHs resulting in the different studies using their own criteria in the analysis of homozygosity. Compared to the era of genotyping microsatellite markers, the advent of high-density single nucleotide polymorphism genotyping arrays has provided an unparalleled opportunity to comprehensively detect these regions in the whole genome in different populations. Several studies have identified ROHs which were associated with complex phenotypes such as schizophrenia, late-onset of Alzheimer's disease and height. Collectively, these studies have conclusively shown the abundance of ROHs larger than 1 Mb in outbred populations. The homozygosity association approach holds great promise in identifying genetic susceptibility loci harboring recessive variants for complex diseases and traits.

Introduction

Human genetic variations are the differences in DNA sequences within the genome of individuals within populations. These variations can take many forms, including single nucleotide variants or substitutions, tandem repeats (short tandem repeats and variable number of tandem repeats), small indels (insertions and deletions of a short DNA sequence), duplications or deletions that change the copy number of a larger segment of a DNA sequence ($\geq$1 kb) i.e. copy number variations (CNVs), and other chromosomal rearrangements such as inversions and translocations (also known as copy-neutral variations) (Nakamura 2009; Frazer et al. 2009; Ku et al. 2010a). The amount of genetic variation in the human genome is more abundant than previously thought, and this has been further corroborated with the findings from whole genome resequencing studies where several million single nucleotide polymorphisms (SNPs) and several hundred thousand indels and structural variants were identified (Wheeler et al. 2008; Bentley et al. 2008; Wang et al. 2008; Kim et al. 2009). In addition to SNPs (Altshuler et al. 2008; Hindorff et al. 2009), other genetic variations have also been found to be associated with various complex diseases and traits (Haberman et al. 2008; Hannan 2010; Wain et al. 2009; Stankiewicz and Lupski 2010).

By comparison, the region of homozygosity (ROH) is not currently classified as a type of genetic variation as there is no consensus on whether it should be classified as one type of 'structural' genetic variation. The reasons for this are two fold: (a) the ROH is not a 'genetic alteration' of the DNA sequence like other genetic variations and, (b) the research on their genome-wide mapping is still relatively new. However, the extent of ROHs varies among individuals and between different populations. In comparison to other types of genetic variations where the inter-population differences have been well documented (International HapMap Consortium 2005, 2007; Jakobsson et al. 2008; Teo et al. 2009), published data has increasingly shown the

C. S. Ku (✉) · N. Naidoo · S. M. Teo
Department of Epidemiology and Public Health,
Centre for Molecular Epidemiology,
Yong Loo Lin School of Medicine,
National University of Singapore, Singapore, Singapore
e-mail: g0700040@nus.edu.sg

Y. Pawitan (✉)
Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden
e-mail: yudi.pawitan@ki.se

inter-individual and inter-population variations in the profiles of homozygosity (Gibson et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; O'Dushlaine et al. 2010).

Research on ROHs has started to gain impetus, as is evidenced by the increasing numbers of publications after the first study by Gibson et al. (2006) reporting its abundance in the human genomes of outbred populations. Further studies have investigated the population genetics aspects of ROHs in healthy individuals (Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b), and also performed association analyses to identify ROHs that are associated with complex diseases and traits in a case–control study design (Lencz et al. 2007; Nalls et al. 2009a; Vine et al. 2009; Yang et al. 2010b).

The aim of this paper is to review the recent progress and to elaborate on the issues and challenges in genome-wide mapping of ROHs in the human genome using high-density SNPs genotyping arrays in normal populations and in disease association studies. We also highlight the findings showing associations between ROHs and complex phenotypes. Finally, we discuss the future directions and the potential applications of ROHs as surrogate markers in identifying recessive loci for complex phenotypes. This approach is also known as 'genome-wide homozygosity association' and could be a promising alternative to finding the 'missing heritability' for complex phenotypes (Manolio et al. 2009). Population genetics and selection pressure on ROHs are briefly discussed, as these topics are beyond the scope of this review paper. Other interesting areas of ROHs research such as studies of homozygosity in inbreeding and isolated populations and findings from animal and plant genetics deserve to be reviewed in a separate paper.

What is a region of homozygosity?

A ROH defines a continuous or uninterrupted stretch of a DNA sequence without heterozygosity in the diploid state, that is in the presence of both copies of the homologous DNA segment. Thus, all the genetic variations, such as SNPs (biallelic marker) or microsatellites (multiallelic marker) within the homologous DNA segments have two identical alleles that create homozygosity (Gibson et al. 2006). The ROH is different from one-copy deletion (or hemizygous deletion), which could also lead to the homozygosity, e.g. in genome-wide SNPs genotyping data. However this is considered as a 'spurious homozygosity' because only one allele of the SNPs is present in the deleted region for one-copy deletions. Thus, the DNA fragments with only the single allele are hybridized on the genotyping array. As a result, the signal intensity of only one allele is measured and subsequently used in genotype calling, and hence it would be incorrectly labeled a homozygote genotype. Therefore, the result of 'homozygosity' is due to the absence of the other allele, instead of 'true homozygosity' where two identical alleles are present (Peiffer et al. 2006). The distinction between 'true homozygosity' as opposed to 'spurious homozygosity' due to one-copy deletion is difficult to determine just by inspection of the genotype data alone. The allelic signal intensity ratio (the relative ratio of the fluorescent signals between two probes/alleles at each SNP) is needed to differentiate between the two types of homozygosity (Peiffer et al. 2006; Wang et al. 2007). Therefore, for studies that used only SNPs genotype data to identify the ROHs, i.e. to screen regions with a minimum consecutive homozygote SNPs, the possibility that some regions are caused by one-copy deletion cannot be firmly excluded, because deletions are also widespread in the human genome (McCarroll et al. 2008; Conrad et al. 2010).

Cytogenetic abnormalities such as uniparental isodisomy can also result in homozygosity where two copies of a single parental homologous DNA segment are inherited from one parent. As such it cannot be distinguished from homozygosity resulting from other factors such as parental consanguinity using the allelic signal intensity ratio as in the case of one-copy deletion. Thus for studies that involved unrelated samples where checking the Mendelian transmission errors in the ROHs is not possible, the possibility of uniparental isodisomy leading to homozygosity cannot be definitively ruled out. Assessing the transmission errors requires data from trios or families. However, the likelihood that a considerable fraction of ROHs will be accounted for by uniparental isodisomy is low given that this cytogenetic abnormality is rare (Curtis 2007).

Currently, there is no consensus or standardized criteria used to define the ROH. However, previous studies have focused on regions $\geq$1 Mb, and thus the true extent of homozygosity in the human genome could be underestimated (Gibson et al. 2006; Li et al. 2006). More recent studies have defined a ROH at a minimum length of 500 kb (Yang et al. 2010b) with the intention of avoiding underestimation of the numbers of regions in the human genome. This is because shorter ROHs are now also thought to be associated with complex phenotypes. However, setting a shorter length for definition will increase the number of false positive signals i.e. increase the sensitivity at the expense of specificity. Therefore, in discovery studies, balancing both the sensitivity and specificity when setting the criteria to identify ROHs is critical.

By focusing only on regions $\geq$500 kb or 1 Mb, the 'noise' introduced by one-copy deletions is likely to be minimal, thus reducing the potential to cause spurious homozygosity. This is because large deletions of $\geq$500 kb are relatively rare in the human genome—as supported by data from high-resolution genome-wide mapping of CNVs

studies (McCarroll et al. 2008; Conrad et al. 2010; Ku et al. 2010b; Park et al. 2010a; Yim et al. 2010). Therefore, a critical issue to be addressed in future homozygosity mapping studies is determining the optimal cutoff of the length of the ROH to be adopted, as this will avoid over-estimating the homozygosity when the length is set too low and which can then be easily confounded by one-copy deletion of hundreds of kilobases or smaller. Although some studies have reduced the cutoff length to 500 kb (Yang et al. 2010b), it is still uncertain whether this new cutoff can readily reflect the true extent of homozygosity in the human genome.

Defining criteria and terminologies

Before the term 'copy number variation (CNV)' was first introduced in 2006 (Freeman et al. 2006), various different terms were used to describe these copy number variable regions such as 'large-scale copy number variants' and 'intermediate-sized variants' (Sebat et al. 2004; Iafrate et al. 2004). To date, various terminologies have also been used to describe the ROHs such as 'extended tracts of homozygosity' (Gibson et al. 2006), 'long contiguous stretches of homozygosity' (Li et al. 2006), 'runs of homozygosity' (Nothnagel et al. 2010; McQuillan et al. 2008), 'autozygosity regions' (Nalls et al. 2009b) and 'homozygosity-by-descent' (Polasek et al. 2010). Different studies have used their own criteria in identifying ROHs with some studies employing more stringent criteria compared to others applying a more liberal definition (Gibson et al. 2006; Li et al. 2006; Nothnagel et al. 2010; McQuillan et al. 2008; Nalls et al. 2009b; Curtis et al. 2008). For example, Curtis et al. (2008) used their own developed software and the criteria of a minimum of 10 consecutive, homozygous SNPs extending over 1 Mb. In comparison, other studies employed the default definition implemented in the 'Runs of homozygosity' function in the PLINK software (http://pngu.mgh.harvard.edu/~purcell/plink/). These criteria are (a) the length of the ROH $\geq$1 Mb, (b) a minimum of 100 SNPs per ROH, and (c) a density of at least 1 SNP per 50 kb (Nothnagel et al. 2010). As all the studies are referring to the same type of 'DNA sequence feature' it is essential to standardize the terminology to be used in describing these regions to avoid confusion.

Polymorphic markers used to detect ROHs

Although long continuous ROHs have been documented a decade ago in reference families from the Centre D'etude Du Polymorphisme Humain (CEPH) (Broman and Weber 1999), no large-scale population-based study had been performed to interrogate the extent of ROHs in the human genome until the first study by Gibson et al. (2006). The recent advances in genome-wide mapping or detection of ROHs have been driven mainly by the availability of highly accurate SNPs databases such as the International HapMap Project, and the technology to genotype several hundred thousand to several million SNPs throughout the human genome (International HapMap Consortium 2005, 2007; Gibbs and Singleton 2006; Ragoussis 2009). The early study in the CEPH families used approximately 8,000 short tandem repeat markers and detected long continuous ROHs. In contrast, subsequent studies have applied SNPs as the polymorphic markers to detect the ROHs (Gibson et al. 2006; Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b). At the single marker level, short tandem repeats are more informative than SNPs because they are multiallelic markers. However, SNPs are more numerous and collectively can yield more information than short tandem repeats and offer a higher resolution compared to other genetic markers—both of which are important to accurately identify the numbers and sizes of ROHs.

Genotyping a large number of SNPs in a microarray platform presents a powerful tool to detect ROHs comprehensively across the whole genome (Gibbs and Singleton 2006; Ragoussis 2009). This also enables investigation into the number, length or size, and location or distribution of the ROHs in the human genome in a more unbiased manner compared to microsatellite markers (Gibson et al. 2006; Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b). The SNPs genotyping platforms also allow studies of the relationship between ROHs and recombination or linkage disequilibrium (LD) patterns, as the SNPs data can be used for haplotype analyses and to calculate the recombination rates (Curtis et al. 2008). The ability to investigate the co-occurrence of ROHs in the areas with extensive LD or low recombination is important in investigating the mechanisms contributing towards the high frequency of ROHs in the human genome.

Genotyping of a sufficiently large number of SNPs is required to accurately detect the ROHs. The study by Gibson et al. (2006) used data from the International HapMap Phase I Project comprising of approximately 1 million SNPs (International HapMap Consortium 2005), whilst other studies have used lower density genotyping arrays ranging from 300,000 to 550,000 SNPs. The importance of having high-density polymorphic markers was shown by Gibson et al. (2006) who found the largest ROH of 17.9 Mb containing 3,922 SNPs from the SNPs data from HapMap Phase I. However, using the data from HapMap Phase II comprising of >3 million SNPs (International HapMap Consortium 2007), a total of 12,778 SNPs were found in the region with 11 heterozygotes. These heterozygotes interrupted the ROH and have divided it into 12 smaller

segments (Gibson et al. 2006). However, it is unclear whether these 11 heterozygotes are genotyping errors or true heterozygotes occurring as a result of recent mutations. Thus, to account for genotyping errors, studies have allowed some missing genotypes and heterozygotes for each ROH to avoid artificially splitting the region (Table 1).

This hints that the sizes of ROHs may be over-estimated in previous studies when using lower density SNPs genotyping arrays. Therefore, the numbers and sizes of ROHs identified by previous studies are likely to be different or altered when higher density SNPs data is available for analysis on the same samples. This also implies that a cautious interpretation should be imposed for ROHs of several megabases for studies using lower resolution SNPs data. A higher density of SNPs is needed for a definitive assessment of ROHs. Although the SNPs genotyping array is an invaluable tool to detect ROHs, it is not without limitations. Similar to CNV detection using SNPs genotyping platforms, the boundaries of the ROHs cannot be determined accurately at a single nucleotide resolution, as accuracy depends on the SNPs resolution. Therefore, like CNVs, the sizes of ROHs could be inflated, i.e. the ROHs detected in previous studies could be smaller than currently estimated. However, there is currently no data supporting this speculation for ROHs as compared to CNVs (McCarroll et al. 2008; Perry et al. 2008).

Methods of detecting ROHs

Several targeted and genome-wide molecular methods are available to detect structural variations such as CNVs (deletions and duplications) and copy-neutral variations (translocations and inversions). However, unlike with structural variations, ROHs cannot be detected with technologies used in molecular genetics such as fluorescence in situ hybridization (FISH) and bacterial artificial chromosome (BAC) clone or oligonucleotide-based comparative genomic hybridization (CGH) arrays (Carson et al. 2006; Feuk et al. 2006; Carter 2007). Furthermore, several new sequencing-based approaches for detecting structural variations such as paired-end sequencing mapping and depth-of-coverage of the sequence read are also unfit to detect ROHs (Korbel et al. 2007; Kidd et al. 2008; Yoon et al. 2009).

The genome-wide mapping of ROHs can only be done using SNPs genotyping arrays or direct sequencing. The whole-genome resequencing or de novo genome assembly using the next or third generation sequencing technologies will offer an almost complete solution to detecting most of the genetic variations including ROHs within the human genome. However, these high-throughput sequencing technologies were not readily available until recently, and the cost is still prohibitively expensive to sequence the whole human genome in a population-based study (Mardis 2008; Metzker 2010). As a result, SNPs genotyping arrays are the main tools for ROH mapping. The SNPs data can be used in two different ways to detect the ROHs. The first approach is to screen the whole genome in a sliding window manner for consecutive SNPs showing homozygotes over a certain length such as 1 Mb, as implemented in PLINK (Purcell et al. 2007). Since this approach only uses genotype data, it is unable to distinguish between true homozygosity and the spurious homozygosity caused by one-copy deletion without further investigations of CNVs in the samples.

This limitation has been overcome by the second approach which relies on the signal intensity data. Two types of signal intensity data are generated by the SNPs genotyping array: (a) the total signal intensity or log R ratio (LRR) and (b) the allelic intensity ratio or B allele frequency (BAF). The combination of LRR and BAF can be used to determine several different states of copy numbers such as homozygous and hemizygous deletions, and one-copy and two-copy duplications, and ROHs as implemented in the PennCNV algorithm. The BAF is needed to differentiate between ROH from normal diploid copies and one-copy deletion (Wang et al. 2007). Figure 1 illustrates the difference in LRR and BAF patterns between ROH and one-copy deletion. For the one-copy deletion, there is a decrease in LRR in addition to the absence of heterozygosity as shown in the BAF panel. Conversely, no reduction in LRR will be seen for ROH, but the absence of heterozygosity is observed. Most of the genome-wide studies of ROHs have used SNPs genotyping arrays. In comparison, the commonly used oligonucleotide-based CGH arrays in detecting CNVs produced only total signal intensity data. This renders them unable to be used for identifying ROHs.

In addition to the most commonly used PLINK software for detecting and analyzing ROHs (Table 1), other methods have also been recently developed for these purposes (Seelow et al. 2009; Browning and Browning 2010; Polasek et al. 2010). The development of powerful and accurate tools or methods for the detection and analysis is a prerequisite for the success of research into ROHs. Furthermore, new algorithms to identify disease-related segments based on homozygosity using case–control data have also been developed. This will enhance studies to identify ROHs that differ between cases and controls, as these regions may contain recessive variants underlying the diseases (Wang et al. 2009). All the ROHs detection methods have their own strengths and limitations with varying rates of false-positive and false-negative results and as such, a combination of methods would be more ideal to minimize these limitations.

Table 1 Summaries of genome-wide association studies of ROHs and complex phenotypes using high-density SNP genotyping arrays

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Schizophrenia (Lencz et al. 2007)	178 cases and 144 controls Affymetrix 500K	Software • Whole-genome homozygosity analysis (WGHA) performed with customized python scripting in the HelixTree environment Criteria • ROH—any window of 100 or more consecutive SNPs that are homozygous, not receiving a heterozygous call • Common ROHs—only those ROHs in which 10 or more subjects share ≥100 identical homozygous calls were retained for further analysis Association analysis • Case–control comparisons of frequency of presence for each common ROH were examined by using χ^2 test	• A total of 339 common ROHs were identified • Schizophrenia cases demonstrated a significantly greater number of common ROHs than controls • 9 ROHs significantly differed in frequency between cases and controls
Bipolar disorder (Vine et al. 2009)	553 cases and 547 controls Affymetrix 500K	This study applied the WGHA approach as demonstrated in the Lencz et al. (2007) study	• A total of 239 common ROHs were identified • The total number of common ROHs did not differ between cases and controls • 7 common ROHs were significant at $p < 0.05$
Late-onset Alzheimer's disease (Nalls et al. 2009a)	837 cases and 550 neurological normal controls Affymetrix 500K	Software • PLINKv1.02 • A sliding window of 50 SNPs, allowing at most 2 missing genotypes and 1 heterozygote call per ROH Criteria • ROH—at least 1 Mb of consecutive homozygous genotypic calls •Minimum SNP density coverage—at least 50 SNPs per megabase Association analysis • 1,090 consensus regions from overlapping ROHs were defined •Each consensus region was found in no less than 10 participants • The consensus ROHs were analyzed using the maxT permutation test algorithm for case/control studies in PLINKv1.02	• One homozygous consensus region in chromosome 8 was found to be significantly overrepresented in cases when compared to controls • The cases presented a slightly higher degree of extended homozygosity when compared with the control group

Table 1 continued

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Height (Yang et al. 2010b)	Discovery study 998 US Caucasian subjects Affymetrix 500K Replication study 8,385 Caucasian subjects from the Framingham Heart Study Affymetrix 500K plus 50K supplemental array	Software • PLINKv1.01 • A sliding window of 5 Mb (minimum 50 SNPs), allowing 5 missing SNPs and 1 heterozygous site per window Criteria • A minimum of 100 consecutive SNPs in a ROH • Minimum length for a ROH, 500 kb • Minimum density in a ROH, 50 kb per SNP • Maximum gap between 2consecutive homozygous SNPs—100 kb Association analysis • Individual ROHs were divided into different ROH groups using the homozyg-group command in the Runs of Homozygosity program • For each ROH group containing >50 subjects—Student's *t* test to compare the adult height of subjects with this ROH group to the height of subjects without this ROH group	Discovery study • 113,910 individual ROHs in 998 subjects • For the association analyses between human adult height and ROHs, 3,322 ROH groups containing more than 50 individual ROHs • 80 ROH groups overlapped with copy number polymorphisms and were excluded from the subsequent association analyses. • One ROH group (ROH 12q21.31) was significantly associated with adult height even after Bonferroni correction Replication study • A significant association with adult height was successfully replicated for the ROH group by FBAT analysis
Colorectal cancer (Spain et al. 2009)	921 cases and 929controls Illumina Infinium Human Hap550 BeadChips	Software • PLINKv1.05 • A sliding window of 50 SNPs, allowing 2% heterozygous SNPs and 5 missing calls in each window Criteria • This study initially analyzed ROHs that were ≥50 SNPs in length • Repeated the analysis using a number of different criteria to define a ROH (≥30 SNPs, ≥40 SNPs, ≥60 SNPs, ≥2 Mb, ≥4 Mb, and ≥10 Mb) Association analysis • Statistical analyses were performed using packages available in R	• No evidence was found for an association between total size of the ROHs in each individual and colorectal cancer • This study calculated the frequencies of cases and controls in which one or more ROHs of ≥4 Mb were detected • 159 of 921 (17%) cases and 142 of 929 (15%) controls had ROHs ($p = 0.14$, Fisher's exact test)
Childhood acute lymphoblastic leukemia (Hosking et al. 2010)	824 cases and 2,398 controls Illumina Infinium Human370 Duo BeadChips	Software • PLINKv1.06 • A sliding window of SNPs across the entire genome, 2% heterozygous SNPs were allowed in each window, 5 missing calls per window Criteria • ROH, ≥75 consecutive SNPs • Only ROHs which occurred in ≥10 persons were retained for analysis Association analysis • Subsequent statistical analyses were performed using packages available in R • Comparison of the distribution of categorical variables was performed using the χ^2 test	• A total of 396 ROHs were identified • Patients and controls showed no significant difference in the average number of ROH • 4 ROHs differed significantly ($p < 0.01$) between cases and controls

Table 1 continued

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Breast and prostate cancer (Enciso-Mora et al. 2010)	Breast cancer 1,183 cases and 1,185 controls Illumina Infinium Human550 Duo BeadChips Prostate cancer 1,177 cases and 1,149 controls Illumina Infinium Human217 and Human 317 BeadChips	Software • PLINK v1.06 • A sliding window of SNPs across the genome, 2% heterozygous SNPs were permitted in each window, 5 missing calls per window Criteria • ROH, ≥80 consecutive SNPs • Only considered ROH that occurred in ≥ 10 individuals Association analysis • Subsequent statistical analyses were performed using packages available in R • Comparison of the distribution of categorical variables was performed using the χ^2 test	• A total of 415 and 426 ROHs were identified in breast cancer and prostate cancer series, respectively • 6 ROHs differed significantly ($p < 0.01$) between breast cancer cases and controls. • 4 ROHs differed significantly ($p < 0.01$) between prostate cancer cases and controls

Different studies have applied different filtering or quality control criteria of the genome-wide SNP data and samples before the data was used for ROH analysis and association studies

Mechanisms generating ROHs

Several mechanisms and factors have been postulated to explain the high frequency of ROHs in the human genome namely, parental consanguinity, uniparental isodisomy and the presence of 'common extended haplotypes'. One of the most common and well established mechanisms leading to ROHs of several megabases is parental consanguinity, in which the offspring inherits chromosomal segments that are identical-by-descent from each parent. Published data has shown that the number of ROHs of several megabases increased markedly in the offspring of consanguineous marriages (Li et al. 2006; Woods et al. 2006) with up to 6% of homozygosity anticipated in the genome of the offspring of first cousin marriages (Broman and Weber 1999). Li et al. (2006) showed that in a family with 4 children from first cousin marriages, multiple ROHs ranging from 3.06 to 53.17 Mb were observed in all the children. Woods et al. (2006) also showed a marked increase in homozygosity levels in individuals with a recessive disease whose parents were first cousins, where 11% of their genomes were homozygous on average. Additionally, the cumulative length of ROHs per genome was found to be larger in two isolated rather than in two more cosmopolitan (non-isolated) European populations (McQuillan et al. 2008). Therefore, when compared to outbred populations, there is an expected increase in the level of homozygosity or number of ROHs in populations where consanguineous marriages are prevalent, as well as in isolated populations where limited random mating or a restricted mate choice has taken place. However, this is unlikely to be the main factor responsible for the high frequency of ROHs in outbred populations in which parental consanguinity is uncommon.

Another widely discussed mechanism is cytogenetic abnormalities such as uniparental disomy, which can be divided into uniparental isodisomy and uniparental heterodisomy. Only uniparental isodisomy can cause homozygosity as the offspring inherits two identical copies of a homologous chromosomal segment from only one parent. As a result, no heterozygosity would be observed in that particular homologous chromosomal segment (Ting et al. 2007). Similarly, this is also an unlikely explanation for the abundance of ROHs reported in the literature; given that uniparental disomies are rare genetic abnormalities that can cause severe and rare genomic disorders when their locations affect imprinted genes. Examples of these disorders are Prader–Willi Syndrome, Angelman Syndrome and Silver–Russell syndrome (Gurrieri and Accadia 2009; Van Buggenhout and Fryns 2009; Abu-Amero et al. 2008). This is further supported by previous studies concluding that the ROHs are not due to genetic abnormalities as no excess apparent deviation from Mendelian transmission was observed. More specifically, transmis-

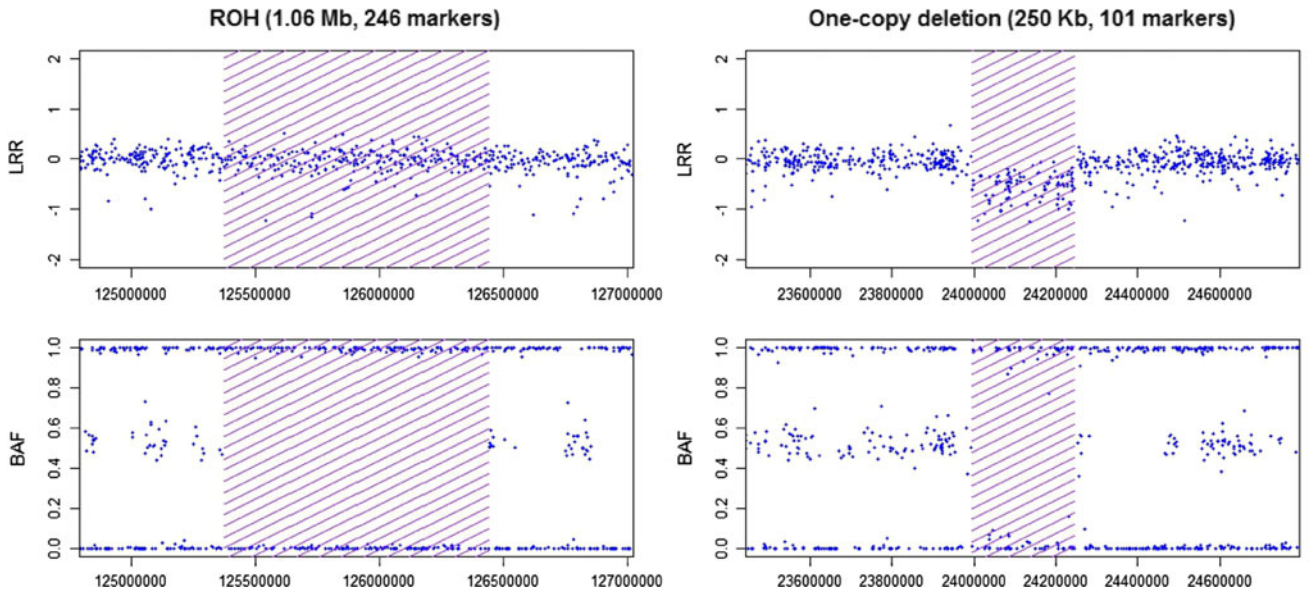


Fig. 1 Plots of the differences in the LRR (Log R Ratio) and BAF (B Allele Frequency) patterns for the ROH (*left panels*) and one-copy deletion (*right panels*) generated from a sample derived from our previous study (Ku et al. 2010b) and genotyped by the Illumina 1 M Beadchip. The ROH and one-copy deletion were detected using the LRR and BAF information by PennCNV algorithm (LRR: total fluorescent intensity signals from both sets of probes/alleles at each SNP, BAF: the relative ratio of the fluorescent signals between two probes/alleles at each SNP) (Wang et al. 2007). The size of the ROH is approximately 1.06 Mb (1,064,933 bases) spanning from 125374832 to 126439764 in chromsome 2. This region contains 246 markers. The size of the one-copy deletion is approximately 250 kb (250,186 bases) spanning from 23994408 to 24244593 in Chromosome 22. This region contains 101 markers. The regions affected by the ROH and one-copy deletion were shaded and the blue dots represent markers in the genotyping array

sion errors occur more rarely in ROHs than would be expected by chance as shown by the observed number of Mendelian transmission errors within a ROH which is less than the expected number (Curtis 2007). Since this study has clearly demonstrated that the ROHs are not usually due to cytogenetic abnormalities, it then indirectly supports the presence of common extended haplotypes as the mechanism contributing toward the high frequency of ROHs in human genomes.

The presence of common extended haplotypes therefore becomes the most likely factor responsible for the high frequency of ROHs which are passed on from both parents to the offspring in the genomes of outbred populations. Data demonstrating the co-occurrence of ROHs in regions with extensive LD and low recombination rates also support the hypothesis of common extended haplotypes in generating homozygosity in the genomes of outbred populations (Gibson et al. 2006; Curtis et al. 2008). A further process believed to be driving the increasing frequency of common extended haplotypes is positive selection. ROHs resulting from common extended haplotypes may be indicative of positive selection pressure of functional importance of these regions. Several methods have been used to quantify the positive selection pressure on ROHs namely, the integrated haplotype score (iHS), Tajima's *D* test and the Fixation index (F_{ST}). Numerous large (several megabases) and common (>25%) ROHs were found to have high values for these metrics indicating the signal for positive selection (Enciso-Mora et al. 2010; Hosking et al. 2010).

Genome-wide mapping of ROHs in the human genome

It was not previously expected that the genomes of outbred populations contain ROHs of several megabases until the first few early reports in 2006 and 2007 (Gibson et al. 2006; Li et al. 2006; Simon-Sanchez et al. 2007). One study found ROHs of >5 Mb in 26 of the 272 unrelated samples assessed (Simon-Sanchez et al. 2007). Similarly, another study performed in Han Chinese also observed the high frequency of ROHs, where 34 of the 515 unrelated individuals contained ROHs ranging from 2.94 to 26.27 Mb (Li et al. 2006). While Gibson et al. (2006) studied the samples from the International HapMap Projects and identified 1,393 ROHs exceeding 1 Mb in 209 unrelated HapMap individuals. Several hundreds of ROHs were found in each of the HapMap populations, and the average number of ROHs (>1 Mb) per individual was found to be lowest in the Yoruba Ibadan Nigerian (YRI) population compared to other populations within the HapMap Phase I Project (Gibson et al. 2006). In addition to demonstrating that ROHs are remarkably common, even in the unrelated individuals from the apparently outbred populations, Gibson et al. (2006) also demonstrated the value of including diverse

populations to examine the differences in ROHs. In the YRI population, the samples have the least number of ROHs per individual. This finding is expected, because the populations of African ancestry are older in human history and hence have more generations and a higher number of recombination events than other populations (recombination occurs during meiosis in each generation). Recombination is one of the important processes to interrupt the long continuous ROHs into smaller segments over the generations. Population differences in ROHs have also been well documented in other studies (Nothnagel et al. 2010).

Each of the previous studies identified a different number of ROHs per individual (Li et al. 2006; Nothnagel et al. 2010; McQuillan et al. 2008; Nalls et al. 2009b; Curtis et al. 2008). These differences are likely reflective of technical and methodological variations such as differing genotyping platforms or SNPs data, differing defining criteria and differing analytical techniques used. Both the genotyping platform and defining criteria can significantly influence the profile of ROHs by way of number, size, cumulative length and genomic distributions. Slight alterations in defining criteria can substantially affect the number of ROHs detected and as a result comparisons between studies are difficult. Therefore, it is critical to develop a set of standardized criteria in identifying ROHs and to establish a database to catalog these regions in the human genome from published studies, similar to other databases developed for SNPs and structural variants (CNVs) such as the dbSNP and Database of Genomic Variants, respectively (Day 2010; Iafrate et al. 2004). This database will enable researchers to quickly compare their results with published data. Consensus on defining the ROHs and the construction of a database to serve as a reference will help in expediting research in ROHs.

LD-pruning of SNPs in mapping of ROHs

The SNPs genotyping data is undoubtedly invaluable for identifying ROHs. However, there is an issue of whether pruning the list of SNPs to remove local LD (i.e. to remove SNPs that are in strong LD) should be done before the data can be used for ROHs. The idea of LD-pruning of SNP data is that the LD between the SNPs can inflate the chance of occurrence of biologically meaningless ROHs. However, there are still uncertainties with regards to the LD-pruning step such as the optimal cutoff of LD (measured by r^2) to be used, although some studies have used the conventional and arbitrary cutoff of $r^2 > 0.8$. More importantly, it is unclear about the quality and performance in terms of sensitivity and specificity for mapping ROHs using LD-pruning SNPs data compared to data without the LD-pruning step. This is an interesting research subject worth pursuing and studies should be done to assess the importance of this LD-pruning step. However, unless significant differences in the sensitivity and specificity are shown using LD-pruning SNP data, the LD-pruning step may not necessarily be needed.

Some of the studies using whole-genome SNPs genotyping arrays have omitted the LD-pruning step before the data was used for mapping ROHs, even though Gibson et al. (2006) used the SNP data from the International HapMap Project where the LD information is readily available. However, others have taken the LD between SNPs into account and used the pairwise LD SNP pruning function in PLINK, with a default value of $r^2 > 0.8$ (Enciso-Mora et al. 2010; Hosking et al. 2010). For example, one study found 370,611 separate tag groups which is a 27.6% reduction of information compared with the original number of SNPs. To account for this, the study adopted a more stringent cutoff of a minimum of 80 consecutive SNPs (instead of 58) to identify ROHs (Enciso-Mora et al. 2010). Similarly Lencz et al. (2007) also took into consideration the LD between the SNPs through setting a more stringent threshold of 100 consecutive SNPs that are homozygous. In comparison, another study removed SNPs in LD with $r^2 < 0.1$ leaving only 30,307 SNPs to form the 'low-LD panel' for some analyses (Spain et al. 2009). Although these studies have taken LD between SNPs into account, it is unclear whether an improvement in sensitivity and specificity was achieved by implementing this LD-pruning step since no evaluation was done to directly compare the differences between the ROHs profile with and without the LD-pruning step. Therefore, the LD-pruning step is conceptually correct; however to warrant this step to be performed in future genome-wide mapping of ROHs, more published data demonstrating its advantages is needed.

Implications on complex diseases and traits

Many novel pathogenic genes or mutations underlying autosomal recessive disorders have been identified through homozygosity mapping. This approach has been shown to be powerful and is particularly useful in investigating autosomal recessive disorders especially in populations with a high prevalence of consanguinity. This is evident from the enormous number of studies identifying causal mutations for autosomal recessive disorders in consanguineous families (Abu Safieh et al. 2010; Harville et al. 2010; Walsh et al. 2010; Pang et al. 2010; Lapunzina et al. 2010; Nicolas et al. 2010; Uz et al. 2010; Iseri et al. 2010; Collin et al. 2010). However, the first study applying the homozygosity association approach at the genome-wide scale for complex diseases only appeared in 2007 (Lencz et al. 2007). Table 1 summarizes the

genome-wide ROH association studies of complex phenotypes using high-density genotyping arrays.

The 'homozygosity analysis' has been shown to be useful for the identification of disease susceptibility genes in both monogenic and complex diseases (Miyazawa et al. 2007; Jiang et al. 2009). The effects of inbreeding or consanguinity and recessive variants or heterozygosity levels on the risk of complex phenotypes (diseases and quantitative traits) have been previously well established (Rudan et al. 2003a, 2003b, 2006; Campbell et al. 2007). A strong linear relationship between the inbreeding coefficient and blood pressure was found and several hundred recessive loci were predicted as contributing to blood pressure variability. Recessive or partially recessive genetic variants account for 10–15% of the total variation in blood pressure (Rudan et al. 2003a). Higher levels of relative heterozygosity were shown to be associated with lower blood pressure and total and low-density lipoprotein cholesterol by measuring genome-wide heterozygosity (Campbell et al. 2007). In addition to quantitative traits, inbreeding was also found to be a significant positive predictor for a number of late-onset complex diseases such as coronary heart diseases, stroke, cancer and asthma (Rudan et al. 2003b). These studies have strongly supported the hypothesis that the genetics of complex phenotypes include a component of recessively acting variants; however, these studies did not directly investigate the associations of complex phenotypes with ROHs detected using polymorphic markers.

Although the information regarding the extent of ROHs in the human genome is still limited compared with SNPs, indels and CNVs, their potential impact on complex diseases and traits could also be significant as other genetic variations. The importance of ROHs to complex phenotypes remains largely unexplored; however, several studies have shown significant differences in ROHs between cases and controls in a genome-wide investigation for schizophrenia (Lencz et al. 2007), late-onset Alzheimer's disease (Nalls et al. 2009a) and height (Yang et al. 2010b). The idea underlying the homozygosity association approach is to uncover recessive variants contributing to complex phenotypes. The success of this approach has been demonstrated in several studies. Nine common ROHs significantly differentiated schizophrenia cases from controls. More interestingly, four of the regions contained or were located near to the genes that are known to be associated with schizophrenia such as *NOS1AP*, *ATF2*, *NSF*, and *PIK3C3* (Lencz et al. 2007). This proof-of-principle study has demonstrated the applications of the whole-genome homozygosity association approach in identifying genetic risk loci for complex phenotypes and it represents an alternative and new avenue in addition to SNPs analysis.

Similarly in a large-scale association study involving 837 late-onset Alzheimer's disease cases and 550 controls, one ROH on chromosome 8 was identified, and three of the genes (*STAR, EIF4EBP1* and *ADRB3*) in the region are biologically plausible candidates (Nalls et al. 2009a). Success was also achieved for complex quantitative traits such as height (Yang et al. 2010b), where strong statistical evidence showing association of one ROH with height was obtained in a total sample size of >10,000 in both the genome-wide discovery and replication studies. The height of individuals with the particular ROH was significantly higher (increased by 3.5 cm) than the individuals without the region. The identification of this ROH added further support to the contribution of recessive loci to adult height variation (Kimura et al. 2008; Xu et al. 2002). Nonetheless, other studies produced negative results, as no evidence of homozygosity was found for bipolar disorder (Vine et al. 2009).

To date, the results showing the association between homozygosity with various cancers are also controversial (Hosking et al. 2010; Assié et al. 2008; Enciso-Mora et al. 2010). For example, two studies investigating the homozygosity in colorectal cancers derived an opposing conclusion which is likely due to the differences between the two studies such as the sample sizes, the density of genotyping platforms and the analysis (Bacolod et al. 2008; Spain et al. 2009). Although studies have found statistically negative results after imposing the stringent Bonferroni correction for multiple-testing, a number of ROHs warrant further investigation as these regions overlapped with biologically plausible genes for the phenotypes. One ROH was found to encompass the gene encoding erythropoietin receptor (*EPOR*) protein. Over-expression of this protein has been documented in acute lymphoblastic leukemia (Hosking et al. 2010).

Many reasons can be speculated for the inconsistencies as to why associations of ROHs were only found in some diseases or studies but not others. This could also indicate that the effects of homozygosity on the risk of complex phenotypes may be disease or trait-dependent, for example some quantitative traits have shown significant variance due to recessive alleles such as systolic blood pressure, total cholesterol and low-density lipoprotein cholesterol. This implies that the effects of homozygosity may be greater in influencing the variation of these traits than others (Campbell et al. 2009). On the other hand, it could also be population-dependent since differences in homozygosity between populations have been documented. Although a number of genome-wide homozygosity association studies have been performed, the optimum study design or analysis methods for assessing the associations or effects of ROHs on the disease risk has not yet been well established. This is, however, vital before breakthrough discoveries can be made in this research area.

The idea for using the homozygosity association approach to dissect the genetics of complex phenotypes is

to reveal the recessive loci that only express their effects (or increase the risk of complex diseases) in the presence of two deleterious recessive alleles, in a recessive disease model. In addition to autosomal recessive disorders, complex diseases can also be affected by recessive variants. The conventional single-SNP analysis approach applied in GWAS may not be statistically powerful enough to identify recessive alleles with small effect sizes and moreover, the recessive model is not usually tested. Until the effect of homozygosity on complex phenotypes is better understood, it is premature to make any conclusions, as the field is still in its infancy compared to association studies between SNPs and CNVs for complex diseases and traits. However, collectively these studies have demonstrated the feasibility of using the homozygosity association approach to identify susceptibility loci for complex phenotypes and have produced encouraging results. This also further underscores the need to further investigate and catalog the extent of ROHs in different populations. Similar to the other genetic variations, ROHs have the potential of becoming the genetic markers in GWAS. In fact, homozygosity mapping has been commonly used to identify the loci for recessive diseases in consanguineous families.

Strengths and shortcomings of genome-wide homozygosity association studies

From the statistical analysis point of view, the advantage of the genome-wide homozygosity association approach is that it suffers lesser penalty from Bonferroni correction for multiple-testing as significantly fewer ROHs are involved compared to the number of SNPs tested in GWAS. Thus, it needs a less stringent *p* value cutoff to declare genome-wide significance. Thus, the genome-wide ROHs association approach has a higher statistical power or requires a fewer number of samples in the studies than the 'conventional GWAS'.

GWAS is an indirect approach that relies on LD to identify the causal variants, thus the results from GWAS are pinpointing genetic loci rather than revealing the causal variants directly (Wang et al. 2005; Hirschhorn and Daly 2005). Similarly in genome-wide homozygosity association studies, one or more ROHs are identified as susceptibility risk loci rather than revealing the actual recessive variants causing the disease. For example, the homozygous consensus region in chromosome 8 was found to be associated with late-onset Alzheimer disease contains seven genes. However, the number of recessive variants within these genes or this region responsible for this 'statistical association signal' and which are functionally important in causing the diseases is unknown (Nalls et al. 2009a). The approaches to be taken from identifying the disease or trait-associated ROHs to locating the functional recessive variants is also unclear. Moreover, the sizes of ROHs are many folds larger than the LD blocks detected by conventional SNP analysis in GWAS, thus making the fine mapping of recessive variants harder. Therefore, the genome-wide association of ROHs, at best, can only pinpoint to a relatively large region harboring as yet to be identified recessive variants.

One common issue and problem in case–control association studies of CNVs and ROHs is how to construct the common CNV and ROH regions in the first place. This step is required to group the individual CNVs or ROHs into a common and discrete region. Similar to CNVs, it is unclear how to partition the individual ROHs into ROH groups so that the frequencies can be used for association analysis. This represents an important analytical challenge in these studies. Genome-wide studies investigating the association of common CNVs with complex phenotypes have so far yielded limited successes (Wellcome Trust Case Control Consortium 2010). As for ROHs, different studies have used their own methods to define ROH groups as no standardized criteria are available. Alternatively this step can be easily performed as the individual ROHs can be divided into different ROH groups by using the 'homozyg-group' command in the 'Runs of Homozygosity' program in PLINK. As a result, each ROH group is actually the overlapping region among all the individual ROHs in the group i.e. the consensus region (the region shared by all overlapping ROHs) (Fig. 2). Using this approach, Yang et al. (2010b) identified 3,322 ROH groups containing more than 50 individual ROHs. While Nalls et al. (2009a) identified 1,090 consensus regions from overlapping ROHs, but each consensus region was found in 10 or more individuals.

Besides identifying the ROH groups for association analysis, attempts were also made to compute other parameters such as the total length of the genome comprised by ROHs (the sum of the length of all ROHs), average length of each ROH (the total length divided by the number of ROHs) and the number of ROHs per individual and

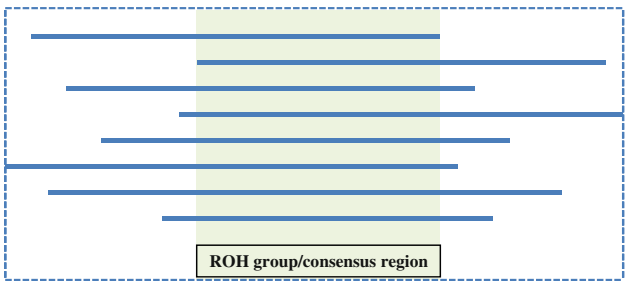


Fig. 2 Schematic diagram illustrating the ROH group or consensus region (*shadowed rectangle*) of several individual ROHs (*blue line*). Only 8 individual ROHs are shown for illustrative purposes with each individual ROH extending in both directions from the consensus region

compare these parameters between cases and controls. Nonetheless, no significant result was observed for late-onset Alzheimer disease (Nalls et al. 2009a). Likewise, no significant difference was found in the average number of ROHs between acute lymphoblastic leukemia, breast and prostate cancers with their controls (Hosking et al. 2010; Enciso-Mora et al. 2010).These analyses may not be very fruitful and have a limited interpretation. Even though significant results were obtained for all the three parameters, the findings are not informative in pointing to specific ROHs that are important to the disease. It can only be concluded that the overall extent of homozygosity is significantly greater in cases compared to controls and thus some recessive variants may be predisposed to the disease risk.

Conclusions

Published data have conclusively demonstrated the high frequency of ROHs in the genomes of outbred populations, and previous studies have also successfully unraveled the associations between ROHs and several complex phenotypes such as schizophrenia, late onset Alzheimer's diseases and height. These studies have shown the promise of the homozygosity association approach in identifying recessive loci for complex phenotypes. However, to what extent this approach contributes toward dissecting the genetics of complex phenotypes is yet to be determined. The analysis of ROHs is now feasible and convenient given the readily available high-density SNPs genotype data and the powerful detection tools such as the PLINK and PennCNV algorithms. Cataloging ROHs in different populations is important, as it lays the foundation for exploring the recessive variants for complex phenotypes.

Currently, the results from GWAS focusing on SNPs analysis alone, explains only a small fraction of the heritability of complex phenotypes (Manolio et al. 2009). Several reasons accounting for the missing heritability have been postulated (Eichler et al. 2010). The missing heritability has challenged the validity of the common-disease common variant (CD/CV) hypothesis (Schork et al. 2009), and has also diverted the research focus to rare variants (Bodmer and Bonilla 2008; Gorlov et al. 2008; Dickson et al. 2010). However, more recent studies have shown that common variants, or more specifically common SNPs, can explain a greater proportion of the heritability than what has been accounted for by GWAS done to date. These SNPs, however, are hidden within the GWAS data, and require larger sample sizes to be discovered (Yang et al. 2010a; Park et al. 2010b). The homozygosity association approach will offer an additional avenue to discovering genetic risk loci that may be missed by the conventional SNPs analysis in GWAS. The homozygosity analysis can be 'easily' performed using the SNPs genotype data and the available detection algorithms, and this is also in line with the ethos of maximizing the information from the GWAS dataset. However several issues and problems still remain as has been discussed.

The power of the homozygosity mapping approach in identifying genes and mutations for autosomal recessive disorders has been previously shown, but currently available data is limited in order to evaluate the success of this approach when applied to complex phenotypes. Hence more studies are needed in the future. Finally we advocate the use of the homozygosity association approach as an additional method of identifying loci harboring recessive variants for complex diseases and traits, which may have been undetected when conventional SNPs analysis was performed alone. The success of this approach has been demonstrated in several complex phenotypes applying the approach. The results so far are encouraging enough to warrant further studies on ROHs to investigate their impacts on complex phenotypes.

Cataloging the ROHs in human genomes and investigating their associations with complex phenotypes should build on the existing GWAS data and these are important areas to pursue in future. The contribution and the role of ROHs in complex phenotypes have been considerably neglected in GWAS; therefore we encourage researchers to explore the associations of ROHs with various phenotypes using their existing SNP data. As the high-density SNPs genotype data have already been generated by several hundred GWAS, the studies of ROHs should be relatively uncomplicated. The availability of these SNP datasets will facilitate the assessment of the roles that ROHs have in complex phenotypes.

References

Abu Safieh L, Aldahmesh MA, Shamseldin H, Hashem M, Shaheen R, Alkuraya H, Al Hazzaa SA, Al-Rajhi A, Alkuraya FS (2010) Clinical and molecular characterisation of Bardet-Biedl syndrome in consanguineous populations: the power of homozygosity mapping. J Med Genet 47:236–241

Abu-Amero S, Monk D, Frost J, Preece M, Stanier P, Moore GE (2008) The genetic aetiology of Silver–Russell syndrome. J Med Genet 45:193–199

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Assié G, LaFramboise T, Platzer P, Eng C (2008) Frequency of germline genomic homozygosity associated with cancer cases. JAMA 299:1437–1445

Bacolod MD, Schemmann GS, Wang S, Shattock R, Giardina SF, Zeng Z, Shia J, Stengel RF, Gerry N, Hoh J, Kirchhoff T, Gold B, Christman MF, Offit K, Gerald WL, Notterman DA, Ott J, Paty PB, Barany F (2008) The signatures of autozygosity among patients with colorectal cancer. Cancer Res 68:2610–2621

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65:1493–1500

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86:526–539

Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, Kolcic I, Weber JL, Hastie ND, Rudan P, Wright AF (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet 16:233–241

Campbell H, Rudan I, Bittles AH, Wright AF (2009) Human population structure, genome autozygosity and human health. Genome Med 1:91

Carson AR, Feuk L, Mohammed M, Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. Hum Genomics 2:403–414

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–S21

Collin RW, Safieh C, Littink KW, Shalev SA, Garzozi HJ, Rizel L, Abbasi AH, Cremers FP, den Hollander AI, Klevering BJ, Ben-Yosef T (2010) Mutations in C2ORF71 cause autosomal-recessive retinitis pigmentosa. Am J Hum Genet 86:783–788

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712

Curtis D (2007) Extended homozygosity is not usually due to cytogenetic abnormality. BMC Genet 8:67

Curtis D, Vine AE, Knight J (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet 72:261–278

Day IN (2010) dbSNP in the detail and copy number complexities. Hum Mutat 31:2–4

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

Enciso-Mora V, Hosking FJ, Houlston RS (2010) Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations. Eur J Hum Genet 18:909–914

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 16:949–961

Gibbs JR, Singleton A (2006) Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS Genet 2:e150

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15:789–795

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112

Gurrieri F, Accadia M (2009) Genetic imprinting: the paradigm of Prader–Willi and Angelman syndromes. Endocr Dev 14:20–28

Haberman Y, Amariglio N, Rechavi G, Eisenberg E (2008) Trinucleotide repeats are prevalent among cancer-related genes. Trends Genet 24:14–18

Hannan AJ (2010) Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet 26:59–65

Harville HM, Held S, Diaz-Font A, Davis EE, Diplas BH, Lewis RA, Borochowitz ZU, Zhou W, Chaki M, MacDonald J, Kayserili H, Beales PL, Katsanis N, Otto E, Hildebrandt F (2010) Identification of 11 novel mutations in eight BBS genes by high-resolution homozygosity mapping. J Med Genet 47:262–267

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hosking FJ, Papaemmanuil E, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JA, Allan JM, Taylor M, Tomlinson IP, Greaves M, Houlston RS (2010) Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk. Blood 115:4472–4477

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium, Frazer KA, Ballinger DG et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Iseri SU, Wyatt AW, Nürnberg G, Kluck C, Nürnberg P, Holder GE, Blair E, Salt A, Ragge NK (2010) Use of genome-wide SNP homozygosity mapping in small pedigrees to identify new mutations in VSX2 causing recessive microphthalmia and a semidominant inner retinal dystrophy. Hum Genet 128:51–60

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jiang H, Orr A, Guernsey DL, Robitaille J, Asselin G, Samuels ME, Dubé MP (2009) Application of homozygosity haplotype analysis to genetic mapping with high-density SNP genotype data. PLoS One 4:e5280

Kidd JM, Cooper GM, Donahue WF et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64

Kim JI, Ju YS, Park H et al (2009) A highly annotated whole genome sequence of a Korean individual. Nature 460:1011–1015

Kimura T, Kobayashi T, Munkhbat B, Oyungerel G, Bilegtsaikhan T, Anar D, Jambaldorj J, Munkhsaikhan S, Munkhtuvshin N, Hayashi H, Oka A, Inoue I, Inoko H (2008) Genome-wide association analysis with selective genotyping identifies candidate loci for adult height at 8q21.13 and 15q22.33–q23 in Mongolians. Hum Genet 123:655–660

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS (2010a) The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet 55:403–415

Ku CS, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A (2010b) Genomic copy number variations in three Southeast Asian populations. Hum Mutat 31:851–857

Lapunzina P, Aglan M, Temtamy S, Caparrós-Martín JA, Valencia M, Letón R, Martínez-Glez V, Elhossini R, Amr K, Vilaboa N, Ruiz-Perez VL (2010) Identification of a frameshift mutation in Osterix in a patient with recessive osteogenesis imperfecta. Am J Hum Genet 87:110–114

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci USA 104:19942–19947

Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MT, Tsai FJ, Chang CF, Wu JY, Chen YT (2006) Long contiguous stretches of homozygosity in the human genome. Hum Mutat 27:1115–1121

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166–1174

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF (2008) Runs of homozygosity in European populations. Am J Hum Genet 83:359–372

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, Hagiwara K (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. Am J Hum Genet 80:1090–1102

Nakamura Y (2009) DNA variations in human and medical genetics: 25 years of my experience. J Hum Genet 54:1–8

Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Launer L, Hardy J, Singleton AB (2009a) Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics 10:183–190

Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB (2009b) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 5:e1000415

Nicolas E, Poitelon Y, Chouery E, Salem N, Levy N, Mégarbané A, Delague V (2010) CAMOS, a nonprogressive, autosomal recessive, congenital cerebellar ataxia, is caused by a mutant zinc-finger protein, ZNF592. Eur J Hum Genet 18:1107–1113

Nothnagel M, Lu TT, Kayser M, Krawczak M (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19:2927–2935

O'Dushlaine CT, Morris D, Moskvina V, Kirov G, Consortium IS, Gill M, Corvin A, Wilson JF, Cavalleri GL (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. Eur J Hum Genet 18:1248–1254

Pang J, Zhang S, Yang P, Hawkins-Lee B, Zhong J, Zhang Y, Ochoa B, Agundez JA, Voelckel MA, Fisher RB, Gu W, Xiong WC, Mei L, She JX, Wang CY (2010) Loss-of-function mutations in HPSE2 cause the autosomal recessive urofacial syndrome. Am J Hum Genet 86:957–962

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Kim H, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo JS (2010a) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42:400–405

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010b) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570–575

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16:1136–1148

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685–695

Polasek O, Hayward C, Bellenguez C, Vitart V, Kolcić I, McQuillan R, Saftić V, Gyllensten U, Wilson JF, Rudan I, Wright AF, Campbell H, Leutenegger AL (2010) Comparative assessment of methods for estimating individual genome-wide homozygosity-by-descent from human genomic data. BMC Genomics 11:139

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a toolset for whole genome association and population based linkage analyses. Am J Hum Genet 81:559–575

Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117–133

Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, Rudan P (2003a) Inbreeding and risk of late onset complex disease. J Med Genet 40:925–932

Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, Rudan P (2003b) Inbreeding and the genetic complexity of human hypertension. Genetics 163:1011–1021

Rudan I, Campbell H, Carothers AD, Hastie ND, Wright AF (2006) Contribution of consanguinuity to polygenic and multifactorial diseases. Nat Genet 38:1224–1225

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212–219

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P (2009) HomozygosityMapper—an interactive approach to homozygosity mapping. Nucleic Acids Res 37:593–599

Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Craw-

ley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16:1–14

Spain SL, Cazier JB, CORGI Consortium, Houlston R, Carvajal-Carmona L, Tomlinson I (2009) Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. Cancer Res 69:7422–7429

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455

Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, Ku CS, Lee EJ, Seielstad M, Chia KS (2009) Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19:2154–2162

Ting JC, Roberson ED, Miller ND, Lysholm-Bernacchi A, Stephan DA, Capone GT, Ruczinski I, Thomas GH, Pevsner J (2007) Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. Hum Mutat 28:1225–1235

Uz E, Alanay Y, Aktas D, Vargel I, Gucer S, Tuncbilek G, von Eggeling F, Yilmaz E, Deren O, Posorski N, Ozdag H, Liehr T, Balci S, Alikasifoglu M, Wollnik B, Akarsu NA (2010) Disruption of ALX1 causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive ALX-related frontonasal dysplasia. Am J Hum Genet 86:789–796

Van Buggenhout G, Fryns JP (2009) Angelman syndrome (AS, MIM 105830). Eur J Hum Genet 17:1367–1373

Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Robinson M, Lawrence J, Anjorin A, Sklar P, Gurling HM, Curtis D (2009) No evidence for excess runs of homozygosity in bipolar disorder. Psychiatr Genet 19:165–170

Wain LV, Armour JA, Tobin MD (2009) Genomic copy number variation, human health, and disease. Lancet 374:340–350

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M (2010) Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet 87:90–94

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17:1665–1674

Wang J, Wang W, Li R et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang S, Haynes C, Barany F, Ott J (2009) Genome-wide autozygosity mapping in human populations. Genet Epidemiol 33:172–180

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME et al (2010) Genome-wide association study of CNVs in 16, 000 cases of eight common diseases and 3, 000 shared controls. Nature 464:713–720

Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, Stern R, Raymond FL, Sandford R, Malik Sharif S, Karbani G, Ahmed M, Bond J, Clayton D, Inglehearn CF (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet 78:889–896

Xu J, Bleecker ER, Jongepier H, Howard TD, Koppelman GH, Postma DS, Meyers DA (2002) Major recessive gene(s) with considerable residual polygenic effect regulating adult height: confirmation of genomewide scan results for chromosomes 6, 9, and 12. Am J Hum Genet 71:646–650

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010a) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yang TL, Guo Y, Zhang LS, Tian Q, Yan H, Papasian CJ, Recker RR, Deng HW (2010b) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. J Clin Endocrinol Metab 95:3777–3782

Yim SH, Kim TM, Hu HJ, Kim JH, Kim BJ, Lee JY, Han BG, Shin SH, Jung SH, Chung YJ (2010) Copy number variations in East-Asian population and their evolutionary and functional implications. Hum Mol Genet 19:1001–1008

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592

Characterising Structural Variation by Means of Next-Generation Sequencing

Chee Seng Ku, *National University of Singapore, Singapore*
Nasheen Naidoo, *National University of Singapore, Singapore*
Shu Mei Teo, *National University of Singapore, Singapore*
Yudi Pawitan, *Karolinska Institutet, Stockholm, Sweden*

Advanced article

Article Contents

Online posting date: 15th February 2011

A new era of copy number variants (CNVs) discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome. Over the past several years, most of the CNV data were generated by microarrays. These methods have several shortcomings, such as the inability to detect copy-neutral variants (e.g. inversions and translocations), limited sensitivity to detect smaller CNVs and poor resolution in determining CNV breakpoints especially with lower resolution microarrays. A paradigm shift in the discovery of copy-neutral variants was attributed to the development of a sequencing-based method known as paired-end mapping. This method was first demonstrated to be powerful in detecting structural variants using next-generation sequencing technologies in 2007. Further studies have also leveraged an important feature of sequencing data, where several hundred million short sequence reads are produced by next-generation sequencers, to detect CNVs based on the abundance or density of the sequence reads aligned to a reference genome. This approach is known as depth-of-coverage. These emerging sequencing-based methods will continue playing an important role in the discovery of structural variants until *de novo* genome assembly becomes more feasible.

ELS subject area: Genetics and Disease

How to cite:
Ku, Chee Seng; Naidoo, Nasheen; Teo, Shu Mei; and Pawitan, Yudi (February 2011) Characterising Structural Variation by Means of Next-Generation Sequencing. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0023399

Introduction

A new era of *copy number variants (CNVs)* discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome (Sebat *et al*., 2004; Iafrate *et al*., 2004). However, these genetic abnormalities were documented decades ago in clinical cytogenetics studies and found to cause various genomic or cytogenetic disorders (Lee *et al*., 2007). The distinguishing feature of the recent studies were that these CNVs were more prevalent in the human genome than expected. These changes in copies number also did not result in any apparent phenotype or disorder and these regions of variable copies were found in the genomes of phenotypically normal individuals (Sebat *et al*., 2004; Iafrate *et al*., 2004). As these submicroscopic (<3–5 Mb) deletions and duplications are beyond the detection limit of traditional cytogenetics tools such as molecular fluorescence *in situ* hybridisation (FISH), these recent discoveries can be credited to the use of whole genome microarray technologies (Carter, 2007). **See also**: Copy Number Variation in the Human Genome; Genetic Variation: Human; Relevance of Copy Number Variation to Human Genetic Disease

Whole Genome Microarray and Sequencing Technologies and Their Progress

The early whole genome microarray studies discovered several hundred CNVs (Sebat *et al*., 2004; Iafrate *et al*., 2004), for example, Sebat *et al*. (2004) detected a total of 221 CNVs in 20 individuals with an average CNV length of 465 Kb. However, it was widely believed that the number of CNVs detected is likely to be underestimated. These

studies used 'low-resolution' microarrays such as ROMA (representational oligonucleotide microarray analysis) containing 85 000 probes with a resolution of approximately one probe for every 35 Kb (Sebat *et al*., 2004) and the BAC-CGH (bacterial artificial chromosome-comparative genomic hybridisation) array with a resolution of approximately one probe for every 1 Mb (Iafrate *et al*., 2004). Furthermore, these studies investigated a small sample size of only tens of individuals which limits the detection of less common CNVs. CNVs smaller than 50–100 Kb will also not be detected as their size is below the resolution limits of these microarrays. Thus, both the sample size and the resolution of microarray are critical factors in determining the discovery of less common and smaller CNVs.

A later study by Tuzun *et al*. (2005) showed that approximately 85% of the 297 identified *structural variants* (139 insertions, 102 deletions and 56 inversions) were not detected by earlier studies. However, this study used a *sequencing*-based method, where the fosmid paired-end sequences were sequenced, instead of microarrays. Many of the structural variants that are being identified using this sequencing-based method are beyond the resolution limit of ROMA and the BAC-CGH microarrays. Inversions are also undetected by microarrays (Tuzun *et al*., 2005; Sebat *et al*., 2004; Iafrate *et al*., 2004). The discovery of many novel structural variants is likely due to the difference between the resolution of sequencing- and microarray-based methods in detecting structural variants.

The contribution of CNVs as a significant source of genetic variation in human populations has since been appreciated despite the limitations using microarrays. This is evident from the enormous amount of interest and efforts generated towards mapping CNVs in different populations (Redon *et al*., 2006; Zogopoulos *et al*., 2007; Wong *et al*., 2007). The first comprehensive mapping of CNVs in the 270 samples from the International HapMap I Project was completed in 2006 (Redon *et al*., 2006). 'Human Genetic Variation' was then recognised as the 'Breakthrough of The Year' in 2007 by the journal *Science*. This was partly accomplished due to the significant progress made in the research of CNVs in addition to the numerous single nucleotide polymorphisms (SNPs) identified by genome-wide association studies for complex phenotypes (Pennisi, 2007). The limitations of ROMA and the BAC-CGH arrays have been overcome in later studies by using higher resolution microarrays and larger sample sizes of several hundred samples (McCarroll *et al*., 2008; Matsuzaki *et al*., 2009; Conrad *et al*., 2010; Park *et al*., 2010; Yim *et al*., 2010; Ku *et al*., 2010). For example, a set of 20 high-resolution oligonucleotide-CGH microarrays comprised of 42 million probes with a median spacing of 56 bases was designed and used by Conrad *et al*. (2010) in mapping CNVs in the HapMap samples (Conrad *et al*., 2010). Other studies have also used the highest resolution SNP microarrays that are commercially available such as the Affymetrix SNP Array 6.0 and the Illumina Human 1M BeadChip (McCarroll *et al*., 2008; Ku *et al*., 2010).

Other types of chromosomal rearrangements, particularly inversions and balanced translocations, have received relatively less attention (Feuk *et al*., 2006; Feuk, 2010; Stankiewicz and Lupski, 2010). Inversions and translocations are also known as 'copy-neutral variants' or 'balanced chromosomal rearrangements' and do not involve changes in copies number (or losses or gains of deoxyribonucleic acid (DNA) sequences). Collectively these copy number and copy-neutral variants are broadly classified as 'structural variants'. The genome-wide mapping or detection of CNVs in different populations has advanced considerably since 2004 and was driven mainly by high-resolution microarray technologies such as oligonucleotide-CGH and SNP microarrays. In contrast, the pace in identifying inversions and translocations in the human genome has been slower as more powerful and effective methods were not available until the advent of *next-generation sequencing (NGS) technologies* (Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010).

Although sequencing-based methods such as *paired-end mapping (PEM)*, which uses cloning and Sanger sequencing methods to sequence the fosmid paired-end sequences, have been shown to be powerful in identifying copy-neutral variants, this method is laborious and expensive (Tuzun *et al*., 2005). Even with the arrival of NGS technologies, PEM has still not as yet been applied in population-based studies (Korbel *et al*., 2007), as opposed to microarrays which are commonly applied to several hundred or thousand samples for CNV detection. However, it is foreseeable in the near future that sequencing-based methods will eventually be routinely and widely applied in large-scale population-based studies when the cost of sequencing becomes more affordable and the challenges in the analysis have been addressed.

The mechanisms that generate structural variants such as nonallelic homologous recombination and non-homologous end joining are beyond the scope of this article (Hastings *et al*., 2009). Similarly, genome-wide detection of CNVs in population-based studies and the population characteristics of CNVs or structural variants, and their associations with various complex diseases or genomic disorders have been reviewed extensively in several excellent review papers (Conrad and Hurles, 2007; McCarroll and Altshuler, 2007). This article will focus on the new and emerging research on structural variants using high-throughput sequencing technologies (Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010; Schadt *et al*., 2010; Gupta, 2008). We also discuss the relative strengths and weaknesses of sequencing-based approaches in comparison to microarrays, and elucidate the potential approaches for a more comprehensive and thorough detection of structural variants in the human genome before *de novo* genome assembly becomes more practical (Li *et al*., 2010a, b; Paszkiewicz and Studholme, 2010).

Microarray-based Methods

Over the past few years, most of the CNV data were generated by CGH and SNP microarrays where fluorescence

signal intensity information was used to detect deletions and duplications. These microarrays are highly accessible and affordable for population-based studies. Additionally, the analysis methods and tools for detecting CNVs using microarray data have been well-developed (Wang *et al*., 2007; Korn *et al*., 2008). This has enabled studies of population characteristics of CNVs in many different populations (McCarroll *et al*., 2008; Matsuzaki *et al*., 2009; Yim *et al*., 2010; Ku *et al*., 2010). However, because of the reliance on the relative or difference in signal intensity compared to a reference in inferring regions with copy number changes, this has hindered microarrays from detecting copy-neutral variants (Carter, 2007). Furthermore, due to the limitations in marker density or resolution of microarrays used in the previous studies, these methods had poor sensitivity to detecting smaller CNVs (<50 Kb) (Redon *et al*., 2006). However, the ability to detect smaller CNVs is critical as they are known to be more numerous than the larger CNVs (Estivill and Armengol, 2007). The accuracy in determining the sizes or breakpoints of CNVs is highly dependent on the resolution of the microarrays as the sizes of CNVs found by previous studies were frequently over-estimated. It is notable that 88% of 1153 CNV loci were smaller than sizes reported in the Database of Genomic Variants and that a reduction of $>50\%$ in size was observed for 76% of the CNV loci (Perry *et al*., 2008).

The latest developments in SNP microarrays such as an increase in marker density and uniformity of distribution in the genome and copy number probes to cover regions with sparse SNPs have improved the sensitivity of microarrays. Nonetheless, these SNP microarrays still lack the sensitivity to detect CNVs smaller than 5–10 Kb even with use of the highest resolution microarrays such as the Illumina Human 1M Beadchip and the Affymetrix SNP Array 6.0 (McCarroll *et al*., 2008; Cooper *et al*., 2008). Although designing a set of high-resolution CGH microarrays comprising tens of millions of probes offers an unprecedented resolution, this method is more costly for several hundred samples (Conrad *et al*., 2010). However, these improvements in microarrays are still unable to detect copy-neutral variants. Thus, developments of other methods that can overcome the limitations of microarrays and simultaneously detect both CNVs and copy-neutral variants are needed.

Sequencing-based Methods

Several previous studies have used sequencing data to detect structural variants. For example, a study by Feuk *et al*. (2005) discovered regions that are inverted between the chimpanzee and human genomes by performing a comparative analysis of their DNA sequence assemblies. This study identified approximately 1600 putative regions of inverted orientation in the genomes (Feuk *et al*., 2005), whereas Khaja *et al*. (2006) identified various types of genetic variants, including structural variants, through comparison of two human assemblies (Khaja *et al*., 2006). However, the paradigm shift in the discovery of copy-neutral variants was attributed to the development of the PEM and concurrent advances in NGS technologies (Korbel *et al*., 2007). The PEM method has also contributed greatly to the discovery of CNVs in the human genome (Wang *et al*., 2008; Ahn *et al*., 2009). **See also**: Comparing the Human and Chimpanzee Genomes; Human Genome Project: Importance in Clinical Genetics; Sequencing the Human Genome: Novel Insights into its Structure and Function

Further studies have also leveraged on an important feature of sequencing data generated by NGS technologies where several hundred million short sequence reads are produced per instrument run to detect CNVs. It is based on the abundance or density of the sequence reads aligned to the reference genome. This approach is known as *depth-of-coverage (DOC)* and is similar to microarray-based methods in that it is also unable to detect copy-neutral variants (Yoon *et al*., 2009). Although *de novo* genome assembly is still developing, the established PEM and DOC methods will continue to play important roles in identifying new structural variants. **Table 1** shows the comparison between microarrays and sequencing-based methods for detecting structural variants.

Paired-end Mapping

Principle

In the PEM method, a library of DNA fragments with a fixed insert size is prepared and both ends of the DNA fragments are sequenced to generate 'paired-end sequences' (the sequences at both ends of the DNA fragments). This sequence information is then aligned against the reference genome. The underlying principle of PEM to detect structural variants is reliant on the discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion. The use of the term 'simple' is to distinguish from other more complex structural variants such as 'everted duplication', 'linked insertion' and 'hanging insertion'. Thus, the terms deletion, insertion and inversion used throughout this paper refer to the 'simple' types unless otherwise specified (Tuzun *et al*., 2005; Korbel *et al*., 2007).

When paired-end sequences that are being aligned to the reference sequence display discordance from the expected insert size or distance, this is an indication of deletion and insertion, whereas discordance in orientation suggests the presence of inversion (i.e. paired-end sequences are incorrectly oriented comparing to the reference genome). Since the insert size of the DNA fragment library is known, when paired-end sequences that align to the reference are substantially shorter than expected, this indicates the presence of insertion. Conversely, a longer than the expected insert size suggests the presence of deletion while other more complicated patterns of discordance when aligning the

Table 1 Comparison between microarrays and sequencing-based methods for detecting structural variants

	Microarrays[a]	PEM[b]	DOC
Principle	Based on the relative or difference in florescence signal intensity compared to a reference (one sample or a set of samples) to infer CNVs	Based on the discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion	Based on the density of sequence reads being aligned to the reference genome to infer CNVs
Ability to detect CNVs	Yes	Yes	Yes
Ability to detect copy-neutral variants	No	Yes	No
Reliably detecting CNVs	Multiple or tens of probes	Multiple discordant pairs	A high density of sequence reads
Application to population-based studies	Commonly applied to several hundred or thousand samples	Has not yet been applied	Has not yet been applied
Sensitivity to detect smaller CNVs e.g. <10 Kb	Generally poor, but depends on the resolution of the microarrays, e.g. a set of oligonucleotide CGH arrays containing 42 million probes has provided an unprecedented resolution	Yes, preparation of several libraries of different insert sizes are able to detect insertions and deletions of varying sizes, but the detection of insertions is limited by the insert sizes	It may not be powerful enough to detect smaller CNVs (related to the strength of DOC signatures and the coverage of the sequencing data or the number of sequence reads)
Sensitivity to detect larger CNVs	Yes, even low resolution BAC clone CGH arrays (with a resolution of approximately one probe for every 1 Mb) have been used to detect CNVs of several hundred kilobases to megabases	Yes, however, the detection of insertions is limited by the insert sizes, thus preparation of fosmid or BAC clone libraries with larger insert sizes are needed for detecting larger insertions	Yes, the DOC signatures will be stronger for larger CNVs
Precision in mapping breakpoints	Generally poor, however, it can be improved by increasing the resolution of microarrays	Good, theoretically the breakpoints can be mapped to a single nucleotide resolution	The precision to map the breakpoints can be improved by increasing the density or coverage of sequence reads
Role in 'discovery' and 'genotyping'	Can be used as an effective method to genotype newly discovered and known CNVs in population-based studies	Powerful for discovery of new structural variants	Discovery of CNVs especially in regions such as segmental duplications where PEM is less effective
Weakness as a result of technology limitation	Generally have poor signal-to-noise ratios for oligonucleotide-CGH and SNP microarrays compared to BAC clone CGH arrays	Short sequence reads are less specific in aligning uniquely to the reference genome especially in segmental duplications	Sequencing biases may lead to certain regions of the genome being over or under-sampled resulting in spurious DOC signatures

(*Continued*)

Table 1 Continued

	Microarrays	PEM	DOC
Scalability of sample throughput by technology	High sample throughput, for example, several hundred samples can be genotyped by SNP arrays per week as evident in genome-wide association studies	Tens of gigabases of sequencing data can be produced per instrument run in several days by NGS technologies, and the sample throughput can be scaled up by 'barcoding' i.e. labelling the samples by barcodes	Tens of gigabases of sequencing data can be produced per instrument run in several days by NGS technologies, and the sample throughput can be scaled up by 'barcoding' i.e. labelling the samples by barcodes
Level of analytical and computational challenges	Lesser, analytical methods for detecting CNVs using microarray data are well-developed	Greater, an emerging and maturing method leveraging on the large amount of NGS data	Greater, an emerging and maturing method leveraging on the large amount of NGS data
Difficulty in sample preparation	Easier in processing the samples for hybridisation on the microarrays	More challenging in preparing sequencing libraries especially clone-based libraries	More challenging in preparing sequencing libraries

[a]Whole genome oligonucleotide-CGH and SNP microarrays.
[b]Paired-end and mate-pair libraries and clone-based libraries (such as fosmid and BAC clones) for PEM.

paired-end sequences provide hints at more complex rearrangements or structural variants (Tuzun *et al*., 2005; Korbel *et al*., 2007; Medvedev *et al*., 2009).

As such, the paired-end sequences are usually classified as 'concordant pairs' or 'discordant pairs' and only the discordant pairs are informative for inferring structural variants. The presence of both concordant and discordant pairs spanning a locus suggests a heterozygote state with respect to the structural variant, for example a deletion occurs only in one homologous chromosome. In addition, usually multiple paired-end sequences are needed to reliably infer if a locus is harbouring a structural variant. The requirement of multiple paired-end sequences spanning a locus to detect structural variants will reduce the number of false-positive signals. It will also minimise the false-negative rate, for example, a heterozygous deletion will be missed by the presence of one concordant pair. However, with multiple paired-end sequences, it is more likely that both the concordant pair and the discordant pair will be observed to detect the heterozygous deletion. As a result, a sufficient amount of sequencing is needed to ensure that there are multiple paired-end sequences spanning across the genome. This also means that a substantial amount of sequencing is needed for the PEM method and thus this method will be more costly using Sanger sequencing compared to NGS technologies (Tuzun *et al*., 2005; Korbel *et al*., 2007; Medvedev *et al*., 2009).

The detection of structural variants using PEM 'signatures' depends on the clustering strategies and criteria used in the analysis, and the results can be varied for the same dataset by applying different strategies and criteria. 'Clustering' refers to steps to group PEM signatures (e.g. several discordant pairs) that support the presence of a structural variant into clusters. As such, clustering will improve reliability in inferring or predicting structural variants and also increase the precision in estimating breakpoints or the sizes of structural variants. The important criteria to be determined in clustering are (a) the minimum number of discordant pairs for a cluster and (b) the number of standard deviations of the insert size to distinguish between concordant and discordant pairs. The strategies and criteria used will then affect the sensitivity and specificity in detecting structural variants (Tuzun *et al*., 2005; Korbel *et al*., 2007; Medvedev *et al*., 2009).

Physical coverage and mate-pair library

'Physical coverage' is important in detecting structural variants using PEM. Physical coverage measures the number of fragments spanning a site and this affects the ability to detect structural variants. It is different from 'sequence coverage' which measures the number of sequence reads that cover a site and this sequence coverage affects the ability to detect single nucleotide variants or point mutations. Thus, physical coverage can be increased by creating a library of larger insert sizes. When preparing a 'shotgun library' using standard methods, the sizes of DNA fragments are usually several hundred bases, with approximately tens of bases on both ends of the DNA fragments sequenced using NGS technologies (Meyerson *et al*., 2010).

However, the insert size can be increased to several kilobases by creating a 'jumping library' or a '*mate-pair library*'. Additional steps are involved in preparing a mate-pair library in comparison to a paired-end library, where both ends of the DNA fragments of several kilobases (e.g.

3 Kb in the Korbel *et al*. (2007) study) were first ligated with biotinylated hairpin adapters. The DNA fragments were then circularised and randomly sheared. The fragments attached to biotinylated hairpin adapters were isolated to form a mate-pair library and then followed by sequencing (Korbel *et al*., 2007). Mate-pair library construction enables sequencing at both ends of longer DNA fragments of several kilobases. The mate-pair library with a larger insert size will increase the physical coverage of the genome. For example, by sequencing 50 bases from both ends of the DNA fragments from a library with a 3-Kb insert size, the physical coverage of the genome is 10-fold higher than that from a library with a 300-bp insert size. However, the sequence coverage is similar between both libraries as only 50 bases of paired-end sequences were generated with regards to the library insert size (Meyerson *et al*., 2010).

Thus the paired-end and mate-pair libraries differ only in the steps of constructing these libraries, as the sequencing and aligning of the paired-end sequences to the reference to detect structural variants follow the same principle. Although creating a mate-pair library increases physical coverage, a larger insert size is less sensitive in detecting smaller structural variants because of the difficulty in tightly controlling the sizes of the DNA fragments in the library. Therefore, depending on the 'tightness' or 'narrowness' of the distribution pattern (standard deviation) of the insert sizes in the library, it can be difficult to distinguish a true PEM signature caused by a small indel (i.e. indel of several or tens of bases) because of the variance in insert sizes in the library. This is because it is not practically possible to generate an exact similar size for each of the DNA fragments when preparing a library (Medvedev *et al*., 2009).

Strengths and weaknesses

In comparison to microarray-based methods, PEM has a higher sensitivity to detect smaller CNVs in addition to identifying copy-neutral variants, and it also has a greater precision in determining the breakpoints or boundaries of structural variants. For example, the PEM method has been applied in a number of whole genome resequencing studies where several thousand structural variants were detected (Wang *et al*., 2008; Ahn *et al*., 2009). Wang *et al*. (2008) identified a total of 2682 structural variants (the majority were CNVs) in the Han Chinese Yan Huang (YH) genome with a median length of approximately half a kilobase. These sizes are much smaller than those identified by microarrays ranging from tens to hundreds of kilobases depending on their resolution (Redon *et al*., 2006; Zogopoulos *et al*., 2007; Wong *et al*., 2007). This has clearly shown the greater sensitivity of PEM to detect smaller structural variants.

Nonetheless, this method could be biased against detection of duplications or insertions. This has been clearly shown in the YH genome, where most of the identified CNVs are deletions, namely 2441 deletions compared to 33 duplications. This is because PEM is unable to detect insertions larger than the insert size of the library. This also reveals the major limitation of PEM with a fixed insert size in detecting insertions (Wang *et al*., 2008). Deletions are easier to be detected because they are identified by a longer than expected insert size when aligned to the reference, whereas detection of insertions is restricted by the insert size. This means that insertions larger than the insert size are beyond the detection range. Therefore, several paired-end and mate-pair libraries with short and long insert sizes will be needed to capture structural variants of varying sizes. This will also nevertheless increase the sequencing costs several fold depending on the number of libraries. For the YH genome, the two paired-end libraries had a small insert size of 135 and 440 bp (Wang *et al*., 2008). Since the bias against detection of insertions is partly due to the small insert size, larger insert sizes of several kilobases should improve the ability to detect more insertions. Indeed, this has been demonstrated by Korbel *et al*. (2007) who prepared libraries of 3 Kb insert size for two individuals and found 1297 structural variants, including 853 deletions, 322 insertions and 122 inversions (Korbel *et al*., 2007). Although the number of deletions is still higher than insertions, it is significantly less biased compared to the numbers detected by Wang *et al*. (2008).

Human Genome Structural Variation Working Group

The PEM method to detect structural variants was first demonstrated by Tuzun *et al*. in 2005 by mapping paired-end sequences data from a human fosmid DNA genomic library. The average insert size of a fosmid library is approximately 40 Kb. However, sequencing of fosmid clones is laborious and costly using Sanger sequencing (Tuzun *et al*., 2005). These limitations have been overcome by NGS technologies which directly sequence the paired-end or mate-pair libraries without the need for cloning steps (Korbel *et al*., 2007). Both of these studies applied the PEM approach to investigate structural variants in the same sample (NA15510) from the International HapMap Project. However, their library insert sizes differed and this has enabled a comparison of the sensitivity between these studies. Korbel *et al*. (2007) were able to confirm 41% of all deletion and inversion events detected by fosmid paired-end sequencing. Additionally, they identified an additional 407 structural variants in NA15510 that had not been previously detected by fosmid paired-end sequencing (Korbel *et al*., 2007; Tuzun *et al*., 2005). This further suggests that several libraries with different insert sizes are needed to increase the sensitivity of PEM. The majority of structural variants detected by PEM were relatively small where approximately 65% were <10 Kb and 30% were <5 Kb (Korbel *et al*., 2007). This represents a significant improvement in resolution over microarrays.

In addition to these studies, a large-scale effort is currently being undertaken by the Human Genome Structural

Variation Working Group to comprehensively map structural variants in phenotypically normal individuals using the PEM approach as demonstrated by Tuzun *et al*. (2005) (Eichler *et al*., 2007). More specifically, the objective is to characterise the pattern of human structural variants at the nucleotide level from a collection of 48 individuals of European, Asian and African ancestry. This project plans to make fosmid clone libraries of approximately 40 Kb insert size from the genomic DNA of 48 unrelated females. These samples have already been genotyped in the HapMap Project. A larger insert size of approximately 150 Kb prepared from BAC clone libraries will also be constructed from 14 unrelated HapMap males. This will aim to provide sequence information on structural variants that are too large to be included in the fosmid libraries, such as those associated with segmental duplications (Eichler *et al*., 2007). As such, both the fosmid and BAC libraries will ensure a comprehensive capture of structural variants of varying sizes across the human genome. A preliminary report was published for eight individuals (Kidd *et al*., 2008).

Depth-of-coverage

Principle, strengths and weaknesses

Depth-of-coverage (DOC) is another method using the NGS data for CNVs detection. As the name implies, this method is based on the depth of coverage of the sequence reads to infer deletions and duplications. The DOC method is enabled by the production of several hundred million short sequence reads per instrument run by NGS technologies. The principle underlying the DOC approach is based on the assumptions that the sequencing process is uniform so that the number of sequence reads mapping to a region follows a Poisson distribution. As such, the number of sequence reads should be proportional to the number of times that a particular region appears in the genome. Therefore, it is expected that a duplicated region will have more reads aligned to it, with the converse true for deletions (Yoon *et al*., 2009; Medvedev *et al*., 2009). However, the assumption that the sequencing process is uniform may not be valid. This is because of the sequencing bias of the NGS technologies which leads to certain regions of the genome being over or under-sampled resulting in spurious signals (Harismendy *et al*., 2009).

Based on the principle of the DOC method, the strength of a DOC signature (i.e. 'gain' or 'loses') is thus directly related to the coverage of the sequencing data (the number of sequence reads) and also to the size of the CNVs. This means that the DOC signatures will be stronger for larger CNVs, and is thus more powerful for detecting larger CNVs compared to PEM. In contrast, unlike PEM, the DOC method cannot detect copy-neutral variants. Moreover, the DOC method may not be powerful enough to identify smaller CNVs (related to the strength of DOC signatures) and it is also limited in defining breakpoints (Medvedev *et al*., 2009). In comparison to microarrays, copies number can only be inferred to four (CN = 4) as the upper boundary for SNP microarray or copy number changes will be denoted as 'gain' or 'loses' for CGH microarrays (McCarroll *et al*., 2008; Wang *et al*., 2008). The DOC method is also more robust and accurate at determining higher copies number.

Merging DOC with PEM

Studies comparing the results between the DOC and PEM methods found that only a small fraction of the CNVs overlap between these methods. Furthermore, the identified CNVs that are specific to the DOC method are more enriched in segmental duplications than the PEM-specific CNVs. This is complementary to the PEM method as it has difficulty detecting structural variants in segmental duplications because the paired-end sequences from these repetitive regions cannot uniquely map to a single site or location in the genome, especially for short sequence reads. In comparison, this problem is less significant for DOC as this method does not rely on uniquely mapping sequence reads to a region to infer CNVs. This suggests that a combination of the methods is ideal to further improve the sensitivity of detection throughout the genome. In fact, both methods have their own advantages and limitations (Yoon *et al*., 2009; Medvedev *et al*., 2009). As discussed earlier, the main assumption of the DOC method may not be valid because of the sequencing biases that cause certain regions to be over or under-sampled. To overcome this limitation, a recent study by Medvedev *et al*. (2010) has developed a method to detect CNVs by supplementing the DOC with the PEM data by integrating both types of sequencing data. Using this integrative method, the discordant pairs will be used to indicate the presence of CNVs for DOC. It has been shown that PEM can improve both the sensitivity and the specificity of the DOC method. Several advantages of integrating the DOC and PEM data have also been demonstrated which addresses some of the limitations of each method when used independently. For example, by using this integrative approach, the size of the variants that can be detected is no longer limited by the insert size of library and this approach is also more robust in detecting variants in segmental duplications (Medvedev *et al*., 2010).

Choosing a Sequencing Platform for PEM and DOC

The applications of high-throughput sequencing technologies that are commercially available and accessible by end-users or researchers for PEM and DOC will be further discussed. It is noteworthy that the development of numerous other sequencing technologies such as single molecule real time (SMRT) sequencing (to be marketed commercially soon) are on the horizon (Schadt *et al*., 2010). Although others such as nanopore sequencing may take several years to become a mature technology (Branton

et al., 2008). In comparison, companies such as Complete Genomics provides a sequencing service rather than selling their sequencing machines to end-users (Drmanac *et al*., 2010). The sequencing technologies that are currently available can be broadly grouped into NGS technologies such as the Roche 454 Genome Sequencer FLX (GS FLX) System, Illumina Genome Analyzer (GA) and Applied Biosystems (ABI) Supported Oligonucleotide Ligation Detection System (SOLiD) and *third generation sequencing (TGS) technologies* such as the HeliScope Single Molecule Sequencer which is now commercially marketed by Helicos Biosciences. **See also**: Next Generation Sequencing Technologies and Their Applications; Whole Genome Resequencing and 1000 Genomes Project

Although Roche 454 GS FLX, Illumina GA and ABI SOLiD are classified as NGS technologies, several features differ substantially between them. They are characterised by the ability of parallel sequencing of a very large number of sequence reads. However, the Roche 454 GS FLX can only generate approximately one million sequence reads per instrument run, in comparison to the Illumina GA and ABI SOLiD where several hundred million sequence reads are produced. Similarly, the HeliScope Single Molecule Sequencer can also produce several hundred million sequence reads (Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010; Li and Wang, 2009). One of the major distinctions between NGS and TGS is that TGS requires no whole genome amplification steps such as emulsion polymerase chain reaction and bridge amplification compared to NGS. Therefore, TGS has the potential to further increase the number of sequence reads or throughput per instrument run than their current capacity. Therefore, the Illumina GA, ABI SOLiD and HeliScope Single Molecule Sequencer provide an advantage for the DOC method that requires a high density of sequence reads to infer CNVs. The specificity of DOC to detect CNVs and the precision to map the breakpoints can be improved by increasing the density or coverage of sequence reads (Yoon *et al*., 2009; Medvedev *et al*., 2009). However, the length of sequence reads produced by Roche 454 GS FLX is on average 400–500 bp, which is substantially longer than that for the other three sequencing technologies which range from 32 to 125 bp (Li and Wang, 2009). Although PEM and DOC methods are targeting large structural variants, the sequence read length produced by Roche 454 GS FLX is better for detecting small indels of several to tens of bases. Moreover, the longer sequence read length of Roche 454 GS FLX may also be more suitable for *de novo* genome assembly before read lengths of several kilobases is generated by future sequencing technologies.

The PEM method, when applying it alone rather than integrated with DOC data, must ensure that the paired-end sequences are uniquely aligned to the reference genome to infer structural variants compared to ambiguous paired-end sequences which align to multiple locations. As such, shorter sequence read lengths may be less specific in aligning against the reference genome especially in repetitive regions such as segmental duplications. Moreover, the number of paired-end sequences is also important as usually multiple discordant pairs are needed to reliably detect structural variants. In terms of preparing the PEM libraries for sequencing, all three NGS technologies are able to generate both paired-end and mate-pair libraries, thus allowing for sequencing of short and longer insert sizes (Robison, 2010; Koboldt *et al*., 2010). Each of the sequencing technologies has its own strengths and weaknesses, and a combination of these technologies in an experiment may be the ideal approach to detecting new structural variants and also to address the systematic biases in sequencing (Harismendy *et al*., 2009).

A Comprehensive Detection of Structural Variants in the Human Genome

Currently no single approach can detect all CNVs or structural variants within a human genome. A combination of different approaches is thus ideal where both microarrays and sequencing-based methods can be utilised for this purpose before *de novo* genome assembly is feasible. In comparison to whole genome resequencing that relies on a reference genome for aligning the sequence reads (Wang *et al*., 2008; Bentley *et al*., 2008; Ahn *et al*., 2009), *de novo* genome assembly will enable a more thorough and comprehensive detection of various genetic variants in the human genome ranging from single nucleotide variants, small indels (insertions and deletions) to large structural variants. Currently *de novo* genome assembly is challenging and less practical because of the short sequence reads generated by NGS technologies especially the Illumina Genome Analyzer and Applied Biosystems SOLiD. However, recent studies have attempted to perform *de novo* human genome assembly using short sequence reads with limited success (Li *et al*., 2010a, b; Paszkiewicz and Studholme, 2010). *De novo* genome assembly will become more feasible with longer sequence read lengths of several to tens of kilobases generated by future sequencing technologies. The number of *de novo* genome assembly studies is anticipated to increase exponentially with the arrival of third generation or single-molecule sequencing technologies in the next few years (Schadt *et al*., 2010; Gupta, 2008; Branton *et al*., 2008).

In anticipation, a recent study has used sequencing and microarray-based strategies to detect various genetic variants which complement the results of the assembly comparison approach used in the HuRef genome (Craig Venter) (Levy *et al*., 2007). This study detected genetic variants by aligning the original Sanger sequence reads generated for the HuRef genome to the reference genome (NCBI build-36 assembly). In addition, high density microarrays were custom-designed to probe the HuRef genome to identify variants in regions where sequencing-based approaches may have difficulties. Thousands of new structural variants (i.e. copy number and copy-neutral variants) were discovered and approximately 1.58% (48.8 Mb) of the HuRef haploid genome consisted of structural variants. In

addition, the study also found biases in each method in detecting these variants. This further justifies the need to combine different methods for a more thorough detection of structural variants (Pang *et al*., 2010).

Conclusions

Microarrays have been widely used in the discovery of CNVs over the last several years. However, with the development of PEM and DOC, this raises the question of whether these sequencing-based methods will eventually replace microarrays in structural variant research. The answer is likely to be a resounding 'yes', but at present the microarrays and sequencing-based methods are proving to be valuable by being complementary to each other in population studies of structural variants. The role of microarrays will likely need to be switched from that of 'discovery' to 'genotyping'. Although sequencing-based methods are more powerful in the discovery of new structural variants, these methods are costly for several hundred or thousand samples especially when several libraries of different insert sizes are needed for PEM. This would limit the number of future studies of population characteristics and disease association. However, the newly discovered and the currently known structural variants can be characterised in population-based studies for investigating their associations with diseases using custom-designed oligonucleotide microarrays. However, this is limited to CNVs which are believed to be in the majority in structural variants. Thus other high-throughput methods to assay newly discovered and known copy-neutral variants need to be developed.

Although the PEM and DOC methods have overcome the major shortcomings of microarrays in detecting structural variants, these methods have their own weaknesses. Nevertheless, these emerging sequencing-based methods will continue to play a role in the discovery of structural variants until *de novo* genome assembly is more feasible (Li *et al*., 2010a, b; Paszkiewicz and Studholme, 2010). *De novo* genome assembly will be more practical with the promise of third generation sequencing technologies to increase the sequence read length to tens of kilobases so that a full human genome can be assembled (Schadt *et al*., 2010; Gupta, 2008; Branton *et al*., 2008). In addition to advancing the knowledge of human genetic variation, these methods are also useful in dissecting somatically acquired rearrangements in cancer genomes (Campbell *et al*., 2008; Stephens *et al*., 2009). Finally, the discovery of various genetic variants including structural variants in the human genome has been greatly accelerated by *1000 Genomes Project* (Genomes Project Consortium, 2010; Sudmant *et al*., 2010).

References

1000 Genomes Project Consortium, Durbin RM, Abecasis GR *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Ahn SM, Kim TH, Lee S *et al*. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research* **19**: 1622–1629.

Bentley DR, Balasubramanian S, Swerdlow HP *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Branton D, Deamer DW, Marziali A *et al*. (2008) The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26**: 1146–1153.

Campbell PJ, Stephens PJ, Pleasance ED *et al*. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40**: 722–729.

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics* **39**: S16–S21.

Conrad DF and Hurles ME (2007) The population genetics of structural variation. *Nature Genetics* **39**: S30–S36.

Conrad DF, Pinto D, Redon R *et al*. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Cooper GM, Zerr T, Kidd JM *et al*. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199–1203.

Drmanac R, Sparks AB, Callow MJ *et al*. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78–81.

Eichler EE, Nickerson DA, Altshuler D *et al*. (2007) Completing the map of human genetic variation. *Nature* **447**: 161–165.

Estivill X and Armengol L (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics* **3**: 1787–1799.

Feuk L (2010) Inversion variants in the human genome: role in disease and genome architecture. *Genome Medicine* **2**: 11.

Feuk L, Carson AR and Scherer SW (2006) Structural variation in the human genome. *Nature Reviews. Genetics* **7**: 85–97.

Feuk L, MacDonald JR, Tang T *et al*. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics* **1**: e56.

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26**: 602–611.

Harismendy O, Ng PC, Strausberg RL *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**: R32.

Hastings PJ, Lupski JR, Rosenberg SM and Ira G (2009) Mechanisms of change in gene copy number. *Nature Reviews. Genetics* **10**: 551–564.

Iafrate AJ, Feuk L, Rivera MN *et al*. (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949–951.

Khaja R, Zhang J, MacDonald JR *et al*. (2006) Genome assembly comparison identifies structural variants in the human genome. *Nature Genetics* **38**: 1413–1418.

Kidd JM, Cooper GM, Donahue WF *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64.

Koboldt DC, Ding L, Mardis ER *et al*. (2010) Challenges of sequencing human genomes. *Briefings in Bioinformatics* **11**: 484–498.

Korbel JO, Urban AE, Affourtit JP *et al*. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Korn JM, Kuruvilla FG, McCarroll SA *et al*. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**: 1253–1260.

Ku CS, Pawitan Y, Sim X *et al*. (2010) Genomic copy number variations in three Southeast Asian populations. *Human Mutation* **31**: 851–857.

Lee C, Iafrate AJ and Brothman AR (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genetics* **39**: S48–S54.

Levy S, Sutton G, Ng PC *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biology* **5**: e254.

Li R, Zhu H, Ruan J *et al*. (2010a) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**: 265–272.

Li Y, Hu Y, Bolund L and Wang J (2010b) State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human Genomics* **4**: 271–277.

Li Y and Wang J (2009) Faster human genome sequencing. *Nature Biotechnology* **27**: 820–821.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.

Matsuzaki H, Wang PH, Hu J *et al*. (2009) High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. *Genome Biology* **10**: R125.

McCarroll SA and Altshuler DM (2007) Copy-number variation and association studies of human disease. *Nature Genetics* **39**: S37–S42.

McCarroll SA, Kuruvilla FG, Korn JM *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166–1174.

Medvedev P, Fiume M, Dzamba M *et al*. (2010) Detecting copy number variation with mated short reads. *Genome Research* September 21 [Epub ahead of print].

Medvedev P, Stanciu M and Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13–S20.

Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews. Genetics* **11**: 31–46.

Meyerson M, Gabriel S and Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews. Genetics* **11**: 685–696.

Pang AW, MacDonald JR, Pinto D *et al*. (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11**: R52.

Park H, Kim JI, Ju YS *et al*. (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nature Genetics* **42**: 400–405.

Paszkiewicz K and Studholme DJ (2010) De novo assembly of short sequence reads. *Briefings in Bioinformatics* **11**: 457–472.

Pennisi E (2007) Breakthrough of the year. *Human Genetic Variation. Science* **318**: 1842–1843.

Perry GH, Ben-Dor A, Tsalenko A *et al*. (2008) The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics* **82**: 685–695.

Redon R, Ishikawa S, Fitch KR *et al*. (2006) Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Robison K (2010) Application of second-generation sequencing to cancer genomics. *Briefings in Bioinformatics* **11**: 524–534.

Schadt EE, Turner S and Kasarskis A (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**: R227–R240.

Sebat J, Lakshmi B, Troge J *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Shendure J and Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

Stankiewicz P and Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annual Review of Medicine* **61**: 437–455.

Stephens PJ, McBride DJ, Lin ML *et al*. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**: 1005–1010.

Sudmant PH, Kitzman JO, Antonacci F *et al*. (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Tuzun E, Sharp AJ and Bailey JA (2005) Fine-scale structural variation of the human genome. *Nature Genetics* **37**: 727–732.

Wang J, Wang W, Li R *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Wang K, Li M, Hadley D *et al*. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**: 1665–1674.

Wong KK, deLeeuw RJ, Dosanjh NS *et al*. (2007) A comprehensive analysis of common copy-number variations in the human genome. *American Journal of Human Genetics* **80**: 91–104.

Yim SH, Kim TM, Hu HJ *et al*. (2010) Copy number variations in East-Asian population and their evolutionary and functional implications. *Human Molecular Genetics* **19**: 1001–1008.

Yoon S, Xuan Z, Makarov V *et al*. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586–1592.

Zogopoulos G, Ha KC, Naqib F *et al*. (2007) Germ-line DNA copy number variation frequencies in a large North American population. *Human Genetics* **122**: 345–353.

Further Reading

Alkan C, Kidd JM, Marques-Bonet T *et al*. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41**: 1061–1067.

Carson AR, Feuk L, Mohammed M and Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. *Human Genomics* **2**: 403–414.

Hormozdiari F, Alkan C, Eichler EE and Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research* **19**: 1270–1278.

Kidd JM, Sampas N, Antonacci F *et al*. (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods* **7**: 365–371.

Wain LV, Armour JA and Tobin MD (2009) Genomic copy number variation, human health, and disease. *Lancet* **374**: 340–350.